

DCBA-DEA: A Monte Carlo Simulation Optimization Approach for Predicting an Accurate Technical Efficiency in Stochastic Environment

Deng Qiang*, Wong Wai Peng

School of Management, University Sains Malaysia, Penang, Malaysia

(Received: February 9, 2014 / Revised: March 27, 2014 / Accepted: Mar 29, 2014)

ABSTRACT

This article describes a 2-in-1 methodology utilizing simulation optimization technique and Data Envelopment Analysis in measuring an accurate efficiency score. Given the high level of stochastic data in real environment, a novel methodology known as Data Collection Budget Allocation-Data Envelopment Analysis (DCBA-DEA) is developed. An example of the method application is shown in banking institutions. In addition to the novel approach presented, this article provides a new insight to the application domain of efficiency measurement as well as the way one conducts efficiency study.

Keywords: Simulation Optimization, Data Envelopment Analysis (DEA), Data Collection Budget Allocation

* Corresponding Author, E-mail: dq11_man112@student.usm.my

1. INTRODUCTION

Data Envelopment Analysis (DEA) has been widely discussed in both methodological and practical areas since the groundbreaking work of Charnes *et al.* (1978). DEA is a powerful mathematical method to measure the relative efficiencies of a set of similar decision-making units (DMUs) which utilize the same inputs to produce the same outputs (Seiford, 1996). DEA is now a standard technique in performance measurement for many areas (e.g., Banker *et al.*, 2013; Cooper, 2013; Ogawa and Ishii, 2002; Okudan and Lin, 2009; Panta *et al.*, 2013; Trinh *et al.*, 2012).

In its name only, DEA is based on the observed data and it implicitly requires the inputs and outputs data should be known exactly. However, in the real environment, there are various sources of uncertainties in data

(e.g., human error, technical malfunction, behavioural bias); this may lead to the stochastic nature of data (Dyson and Shale, 2010). Moreover, if the collected data are not representative or missing, the efficiency analysis may not be able to be carried out completely, and the resulting efficiency scores will be erroneous and misleading. Hence, the effects and characteristics of the data is vital source of impact that engenders confidence in the results of the efficiency analysis. In order to address this issue, DEA has been enhanced to handle data uncertainties, in what being termed as stochastic DEA. Various forms of stochastic DEA have been discussed, e.g., imprecise DEA (IDEA), chance constrained DEA (CC DEA), bootstrapping, and Monte Carlo simulation.

IDEA can be applied for bounded and ordinal data; this approach is effective only when selecting the extreme values over the ranges whereas the extent of the

ranges may impact the different degrees of uncertainty in the data. Despotis and Smirlis (2002) developed an interval efficiency by assuming that the efficiency of a DMU is an interval between the pessimistic and optimistic condition measures. Although this method can determine the lower bound and upper bound of the efficiency score, the bound is too wide to make reasonable evaluations to each DMU, hence this approach is more suitable for sensitivity analysis rather than to measuring stochastic efficiency (Zhu, 2003).

In CC DEA, the data of inputs are deterministic while the outputs are stochastic with a distribution; the amputation of uncertainty is determined by allowing a DMU to lie outside of the efficient frontier with a certain probability. This approach is intractable because it can only incorporate the uncertainty in outputs rather both inputs and outputs (Cooper *et al.*, 1996; Olesen, 2006; Olesen and Petersen, 1995).

Bootstrapping is a method of measuring the uncertainty in DEA efficiency through generating the confidence intervals; it assumes that the current data is of the one sample set of the population which has an infinite of units by the process of data generating. This approach is able to give a specific estimation to the uncertainty of efficiency in DEA. However, as for the examined inputs and outputs, it can not load any specific information about data uncertainty. Additionally, it is unable to handle the very real case of any observed values being more uncertain than others due to the homogeneity assumption on efficiency (Dyson and Shale, 2010; Fried *et al.*, 2008; Simar and Wilson, 1998).

Monte Carlo simulation has been applied in DEA to handle uncertainty in data by many researchers (e.g., Wong *et al.*, 2008; Kao and Liu, 2009). The studies describe the uncertainty of the data with certain distributions by employing simulation method, to compare the techniques or to examine the accuracy of the efficiency scores. The challenge lies in the process of distribution construction which has the potential of computational greediness. It means that only applying Monte Carlo simulation will waste much computation time on the non-critical alternatives. This is because in the conventional simulation process, each process of collecting data has been set equally with the same number of replications. Wong *et al.* (2011) has suggested the method of optimal computing budget algorithm (OCBA), this can reduce the total computation time effectively for collecting the simulation data. Their works have shown the great potential of soft computing approach in improving the accuracy of DEA efficiency.

Apart from considering which technique to handle the uncertainties in data, it is also important to consider that what activities that needs to be done in carrying out an efficiency measurement study. Logically, before one employs a technique to measure efficiency, it needs to have a set of data for that analysis purpose. This means, one needs to 'collect the data.' With regards to this, let's use the lay man term 'data collection' to denote this ac-

tivity. In the realm of efficiency measurement, *data collection* is the utmost important, yet most time consuming step (e.g., various kinds of *inputs* and *outputs* data for each DMU have to be collected). Here, we need to address that "How should we allocate the computation budget for simulation data collection?"

Therefore, we aim to springboard from this paper, an alternative approach for efficiency measurement. This can be viewed as an upcoming generation of stochastic DEA, which, instead of concentrating on the downstream of the technique (i.e., how to calculate the end results, i.e., efficiency), here we looked at the upstream modus operandi that is the data collection part. This data collection part obtains the data, which eventually feeds into generating the end outputs, i.e., efficiency. This in an innovative approach as it stimulates the thinking of 'out of the box.' In addition, the final result of the proposed technique is the improvement in the accuracy of the efficiency. This idea utilized simulation optimization technique, and this method is known as Data Collection Budget Allocation Data Envelopment Analysis (DCBA-DEA). The paper will proceed in the following manner. First, we will explain about DCBA-DEA and the technical aspect of it. Then, we show a simple illustration on the application of DCBA-DEA. And finally, this paper culminates with the important insights that can be observed through this idea.

2. THE PROPOSED IDEA

Originated from the basic DEA model, the DCBA-DEA aims to provide an accurate measurement of the efficiency when the data are uncertain through the allocation of data collection effort. It is a 2-in-1 approach, where the first step is to obtain the efficiency scores, and then, followed by step two, that is to improve the accuracy of the efficiency scores. This approach can be used to handle uncertainties in data. First, we provide a short description of DEA (Cooper *et al.*, 2000; Thanassoulis, 2001).

2.1 Step 1: DEA-Based Efficiency Measurement

Assume that there is data on S inputs and R outputs for each J DMU. Let K denotes the set of combined inputs/outputs, i.e., $K = S \cup R$. D as the total number of inputs/outputs, $K = \{1, \dots, D\}$. For the j th DMU, these are represented by the vectors x_j and y_j , respectively. Let $\mathbf{X}_D = (x_{kj})_{k \in K, j \in J}$, where x_{kj} represents k th input/output for DMU j . If $k \in S$, then x_{kj} is an input; otherwise if $k \in R$, then x_{kj} is an output. To measure the efficiency for each DMU, we calculate a ratio of all inputs/outputs, such as $(u_{k,k \in R} x_{kj, k \in R} / u_{k,k \in S} x_{kj, k \in S})$ where u is a $K \times 1$ vector of weights; i.e., if $k \in R$, then u_k is an output weight; otherwise if $k \in S$, then u_k is an input weight. To select optimal weights we specify the following mathematical programming problem:

$$\begin{aligned} & \min \left(u_{k,k \in R}, x_{kj,k \in R} / u_{k,k \in S} x_{kj,k \in S} \right), \\ & \left(u_{k,k \in R}, x_{kj,k \in R} / u_{k,k \in S} x_{kj,k \in S} \right) \leq 1, j \in J, u_k \geq 0 \end{aligned} \quad (1)$$

The above formulation has a problem of infinite solutions, by imposing a constraint $(u_{k,k \in S} x_{kj,k \in S}) = 1$, this will lead to the equivalent envelopment form as follows (2) after transformation using the duality in linear programming.

$$\begin{aligned} \theta(\mathbf{X}_D) = \min \theta \\ \text{s.t. } \sum_{j \in J} \lambda_j x_{sj} &\leq \theta x_{sj_o} & s \in S \\ \sum_{j \in J} \lambda_j x_{rj} &\geq x_{rj_o} & r \in R \\ \lambda_j &\geq 0, & j \in J \end{aligned} \quad (2)$$

Where $\theta(\mathbf{X}_D)$ is a scalar representing the value of the efficiency for the j_o th DMU which will range between 0 and 1. λ_j is a vector of $J \times 1$ transformed weights (decision variables) of the inputs/outputs that optimize the efficiency score of DMU j_o .

The above DEA model (2), also known as the constant-return-to-scale model, has to be solved J times, once for each DMU in the sample. In order to calculate efficiency for a group of comparable DMUs (e.g., inefficient DMU is only compared against DMU of similar size), and therefore provides the basis for measuring economies of scale within the DEA concept, a convexity constraint $\sum_{j \in J} \lambda_j = 1$ is added. This is called the variable-return-to-scale model. The convexity constraint determines how closely the production frontier envelops the observed input/output combinations.

The issue in this step is that it does not cater for uncertainties in the data. For example, if one does not have the full set of data, then the respective DMU or the related data category will have to be omitted in the analysis (this defeats the overall purpose of the study). An alternative measure, one may resort to guessing the data; in which, this will render the efficiency score unreliable at the end. Using averages of the data (or mean) may also pose a problem to the accuracy of the data if the sample size is too small, as this may not fulfil the central tendency theorem, i.e., the sample means will be far away from the true mean values.

In view of this, we aim to address the uncertainties of data in efficiency measurement by using the simulation optimization concept, i.e., in our step 2.

2.2 Step 2: Enhancement for Accuracy – DCBA

“How to allocate the budget effectively in order to obtain an accurate efficiency score” is the main principle in this step. The distribution of the efficiency score, which ultimately identifies the tendency of where the true efficiency values normally lies, are determined by

the amount of data collected on the inputs/outputs. In reality, data collection is expensive and subject to budget constraint. In this respect, if the data collection effort is done naively, the entire efficiency study will be jeopardized. For example, if the collected data are not representative, then the resulting efficiencies will be erroneous and misleading. Hence, the issue of how to allocate the budget effectively as to which type of data to collect and how many data to collect are therefore of prime importance in efficiency study.

2.2.1 Conceptualization based on simulation optimization

The issue of “How to allocate the budget effectively?” has been addressed in the field of simulation optimization, where OCBA algorithms has been developed to determine the simulation replications to be allocated to each simulation model in order to identify the best design using the least amount of computing budget (Chen and Lee, 2010; Lee *et al.*, 2013). Here we employed the concept of this technique to develop our DCBA model. From the lessons learned in OCBA, we know that if we naively allocate data collection effort fairly, we may have wasted the effort in collecting data that are not sensitive to efficiency and the accuracy of the estimated efficiency may not improve. Therefore, with this aim, we want to find the best data collection plan (*design*) in terms of best use (*allocation*) of resources (*budget*) that gives the most accurate efficiency score.

The DCBA model

$$\begin{aligned} \min F(\mathbf{n}) &= E \left[(\tilde{\theta}(\mathbf{X}') - \theta(\mathbf{X}))^2 \right] \\ \text{s.t. } \sum_{i \in K} n_i &= N \end{aligned} \quad (3)$$

The above is our DCBA model derived from the OCBA concept. The accuracy of the efficiency is measured through mean square error (MSE). The objective function in the above model, $F(\mathbf{n})$ is defined as the MSE of the efficiency score for allocation design \mathbf{n} where \mathbf{X}' is the belief of the inputs/outputs after additional data are collected following the allocation design \mathbf{n} ; N is the total computing budget (data points) for allocating to k variables (the number of stochastic inputs and outputs). Note that $\theta(\mathbf{X})$ is the efficiency score computed using (2) (i.e., $\theta(\mathbf{X}_D)$), for simplicity, we discard the notation D from \mathbf{X}_D and $\tilde{\theta}(\mathbf{X}')$ represents the belief for the true efficiency.

Note that the model cannot be solved directly, as the solution is not in closed-form formula. Inherently, the above problem is non-convex discrete optimization problem where there is no good structures exist for us to develop simple efficient algorithm. One way to estimate $F(\mathbf{n})$ is to derive \mathbf{X}' through the Bayesian framework. Through the quantification of \mathbf{X}' using Bayesian, F can be estimated for a given value of \mathbf{n} as follows.

$$F(n) \approx \frac{1}{M} \sum_{i=1}^M (\theta(\hat{X}_{[i]}) - \theta(\mathbf{X}))^2 \quad (4)$$

Where $\hat{X}_{[i]}$ is the realization of the inputs/outputs \mathbf{X}' in the replication i of the Monte Carlo run for allocation design \mathbf{n} and M is the cardinality of random data set. This method thus is a *simulation based* technique. Recall that an allocation design is given as $\mathbf{n} = [n_i]_{i \in K}$, K is the combined inputs and outputs, D is the total number of inputs/outputs, N is the total number of data collections, and $K = \{1, \dots, D\}$. As the search space is discrete and very huge, e.g., for $D = 15$, $N = 150$, the possible number of \mathbf{n} (the allocation data collection plan) can be as high as 6.60×10^{19} . If we evaluate every design in the search space, assume each design needs 2 seconds to be searched and 3 seconds to be evaluated, all designs may need 3.3×10^{20} seconds to finish, that is to say, it will cost 1.04×10^{13} years to finish.

In order to reduce the time for searching the design, two search-based methods have been presented by Wong *et al.* (2011). The first is a sequential approach to collecting data, whereby a hill-climbing technique is used to gradually improve the solution and moves towards N starting from an initial value of zeros for all n_k . The second technique is a non-sequential data collection approach whereby a metaheuristic technique (genetic algorithm) is employed to find the optimal solution from a given pool of possible solutions, i.e., starting from the set of possible n_k values that formed N . For more details on how to enhance the model, readers are advised to refer to Wong *et al.* (2011).

2.2.2 Allocation of data collection plan using simulation optimization technique

In order to get the best design (data collection plan), we need to evaluate the designs after the searching design process. While in this design evaluation process, there still exist time wastes if we equally allocate the computing budget for each selected design. Hence, we apply OCBA (Chen and Lee, 2010), a simulation optimization technique in evaluating and choosing the best design with smallest MSE. Next we illustrate on how OCBA helps to reduce the simulation time in the process of design evaluation, as it can intellectually choose to spend much more time and effort in evaluating a design which has potential to get lower MSE. Here, we select the non-sequential data collection plan as it is much simpler for explanation purpose.

To run the OCBA mode, additional notations required are as follows:

- T_{n_i} = number of simulation replications allocated to design \mathbf{n}_i ;
- \bar{B}_{n_i} = sample mean of the MSE for design \mathbf{n}_i ;
- E_{n_i} = variance of the MSE for design \mathbf{n}_i ;
- m = number of top designs to be selected; and
- H = total simulation runs (or computing budget, i.e., used to evaluate the designs).

We limit the size of \mathbf{n} to l in each simulation run. From literature of OCBA (Chen and Lee, 2010), the computing allocation budget (or simulation runs) for each design can be determined through the relationship (5) below.

$$\frac{T_{n_i}^{t+1}}{T_{n_j}^{t+1}} = \left(\frac{E_{n_i} / G_{n_i}}{E_{n_j} / G_{n_j}} \right), \quad i, j \in \{1, 2, \dots, l\} \text{ and } i \neq j \quad (5)$$

Where $G_{n_i} = \bar{B}_{n_i} - (\bar{B}_{n_{i,m}} + \bar{B}_{n_{i,m+1}}) / 2$, $\bar{B}_{n_{i,m}}$ and $\bar{B}_{n_{i,m+1}}$ are the sample means of the MSE for design \mathbf{n}_i in the m and $m+1$ top number of designs selected, respectively.

We first simulate all l designs with t_0 replications. As simulation proceeds, the sample means and sample variances of each design are computed from the data already collected up to that stage. The simulation budget is then increased by Δ and Eq. (5) is applied to determine the new simulation runs allocation. Further simulation replications are then performed based on the allocation and the procedure is repeated until the total runs H is exhausted. The following shows the procedure of OCBA-m allocation.

Step 1: *Initialize*. Set $t = 1$ and perform t_0 simulation replications for all designs.

Set $T_{n_1}^t = T_{n_2}^t = \dots = T_{n_l}^t = t_0$.

Step 2: *Update*. Calculate \bar{B}_{n_i} , E_{n_i} and G_{n_i} for $i = 1, \dots, l$.

Step 3: *Allocate*. Increase the computing budget by Δ and calculate the new budget allocation $T_{n_1}^{t+1}$, $T_{n_2}^{t+1}$, \dots , $T_{n_l}^{t+1}$ according to (5).

Step 4: *Simulate*. Perform additional $T_{n_i}^{t+1} - T_{n_i}^t$ simulations for design \mathbf{n}_i for $i = 1, \dots, l$.

Step 5: *Termination*. If $\sum_{i=1}^l T_{n_i}^t < H$, set $t \leftarrow t + 1$ and return to step 2; otherwise, stop.

To summarize the 2-in-1 (DCBA-DEA) approach is used to address efficiency measurement in the presence of uncertainties of data, whereby the first step is to measure the efficiency, followed by second step, which is the budget allocation procedure for data collection to improve the accuracy of the efficiencies.

3. NUMERICAL EXAMPLE

Consider an example in Despotis and Smirlis (2002). There are five DMUs using two inputs to produce two outputs, both inputs and outputs are interval data. The original data are shown in Table 1 (in this paper, we use $[L, U]$ to indicate the range, where L represents lower bound and U means upper bound). To conform to the formulation of this paper, we add the average for each interval data.

Table 1. The example data of five decision-making units (DMUs)

DMU	Input1		Input2		Output1		Output2	
	Ave	Range	Ave	Range	Ave	Range	Ave	Range
A	30	[25, 35]	17.5	[16, 19]	11	[7, 15]	5.5	[4, 7]
B	21	[19, 23]	22.5	[21, 24]	15	[13, 17]	3.7	[3.5, 3.9]
C	16	[15, 17]	11.3	[10, 12.6]	22.8	[20, 25.6]	7.9	[6, 9.8]
D	37	[35, 39]	18.2	[16, 20.4]	9	[8, 10]	6.1	[5, 7.2]
E	25	[22, 28]	15.3	[14, 16.6]	13	[12, 14]	7.4	[7, 7.8]

Table 2. The result of different approaches

DMU	Despotis and Smirlis (2002)		DCBA-DEA						Accuracy improvement ($\frac{A-B}{A} \times 100\%$)	
	Efficiency range (A)	Mean efficiency	Data collection design				MSE	SD		Efficiency range (B)
			n1	n2	n3	n4				
A	[0.224, 1]	0.542	5	9	9	7	1.95E-05	0.015	[0.515, 0.572]	92.65
B	[0.227, 1]	0.763	15	13	0	2	1.79E-04	0.0374	[0.705, 0.808]	86.68
C	[0.823, 1]	1	12	5	11	2	7.22E-23	0.000	[1, 1]	100.00
D	[0.445, 0.907]	0.691	1	22	5	2	1.80E-05	0.016	[0.666, 0.714]	89.61
E	[1, 1]	1	1	2	10	17	6.09E-29	0.000	[1, 1]	0.00

DMU: decision-making unit, DCBA-DEA: Data Collection Budget Allocation Data Envelopment Analysis, MSE: mean square error, SD: standard deviation.

In order to run the simulations, we set the parameters in DCBA model as follows:

- $D = 4$, which represents the total number of inputs and outputs;
- $N = 30$, which means the total number of data collections;
- $t_0 = 10$, which represents the initial simulation replications;
- $l = 20$, which represents the number of designs been evaluated for each simulation run;
- $H = 200$, which means the total simulation replications or computing budget;
- $\Delta = 20$, which is the computing budget increasing per each MSE evaluation.

In this sample example, the number of possible designs is only 5456, so we apply the DCBA-DEA without the search-based method. We code them in MATLAB 7.0 and ran them in an operation system of Windows7 (CPU 1.6 GHz, RAM 2 GB). After the simulation, a feasible solution $\beta = [n1, n2, n3, n4]$ represents the additional number of data to be collected for the four variables (input1, input2, output1, output2), respectively. Meanwhile, to avoid the computing chanciness to impact the simulation result, the entire process is executed 200 times to obtain the standard deviations and confidence intervals of the efficiency scores. The left part of Table 2 shows the results from the Despotis and Smirlis (2002), and the right part of Table 2 shows the result by DCBA-DEA.

Despotis and Smirlis (2002) only calculated the bounds of the efficiency score for this example, and their efficiency ranges are too wide to make one DMU classify from another. Our result shows narrower ranges of efficiency score, the average improvement in the accuracy of efficiency ranges is above 85%. More importantly, our method not only can get the statistic information for the efficiency, but it also can find the best simulation design with time reduction and direct to the real data collection. As a result, the efficiency scores for DMUs A, B, and D lie in the ranges of [0.515, 0.572], [0.705, 0.808], and [0.666, 0.714], which have the accuracy improvement in narrow ranges with 92.65%, 86.68%, and 89.61%, respectively. Take DMU A as example, under the assumption of normal distribution, Figure 1

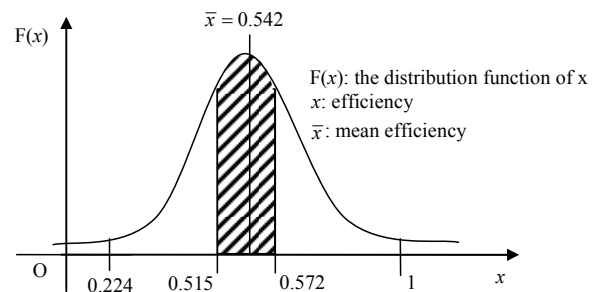


Figure 1. The statistics of efficiency for decision-making unit (DMU) A.

shows the statistics information of efficiency for DMU A, the shadowed part visually represents the range of efficiency score by DCBA-DEA is much more narrower compared with Despotis and Smirlis (2002).

Notably, the efficiency scores for DMUs C and E are unity, despite that the data are stochastic. This is because the whole ranges of the interval data of these two DMUs lie on the efficient frontier. Meanwhile, the best data collection design with MSE for simulation have been presented in Table 2, we find that the narrower is the efficiency range, the smaller is the MSE. This indicates that the mean square error, which has been chosen to be a numerical measure of accuracy level for efficiency score in the DCBA-DEA mode, is verified as a critical factor of reference value for the level of accuracy. We also find that the DMU with narrower efficiency range has got lower standard deviation; in the statistics, the standard deviation represents the closeness of simulation efficiency to the mean efficiency and hence it can express the stability of the simulation process. That is to say, the lower standard deviation represents the higher reliability of the result. In this example, even for the DMU B that has the highest standard deviation with the value of 0.0374, is only take 4% of its mean efficiency. Nevertheless, this example only has five DMUs with four variables; all the possible designs can be evaluated without the search-based method. Next, we will illustrate the more complex problem in real environment.

4. AN ILLUSTRATION ON BANKING INSTITUTIONS

Efficiency measurement of banks is chosen as the scope of study. The reason of choosing this field is that, efficiency measurement of banks often has to compile yearly data; and the problem with this is most data are not complete. Sources of this uncertainty are commonly due to the human error in managing data and also the culmination effects of change in government policy, e.g., merging of banks, which lead to difficulties in data separation and integration. Conventionally, when there are observations of multiple years' data, the averages across the years are calculated to represent the tendency of the data and these averages are used to measure the efficiency. In this section, we will use our method to analyze the data of 25 Taiwan commercial banks during 1997–2001 from Kao and Liu (2009). Table 3 shows the five-year data with average, lower bound and upper bound for each bank, where the monetary amounts are billion New Taiwan dollars (100 NTD \approx 3 USD). This data has 25 DMUs with 6 variables; therefore, it has a huge search space for all possible data collection design even for small N . In order to simulate the data sufficiently, we set N as 300 and add the genetic algorithm to help to search the best data collection design with smallest MSE. Next we will compare our results with the

results from conventional method and the simulation approach by Kao and Liu (2009).

In conventional studies, when there have multiple observations in periods, normally the average data are used to measure the efficiency. The second column of Table 4 shows the efficiency scores of 25 banks from the average data. There are five banks (banks 4, 11, 13, 14, and 23) are efficient. Banks 7 and 8 have the smallest efficiency scores of 0.6291 and 0.6534, respectively. The simple method only use average data, this can make the measurement of efficiency easier and gives a general view of each bank's performance in efficiency; while it may produce erroneous results. The method of interval efficiency has been discussed in the numerical example and the efficiency may be too wide to draw conclusions, so here we do not list the result of interval efficiency, instead, we put the result of a simulation method from Kao and Liu (2009), the middle part of the Table 4 shows the result of their approach, the rights part of the Table 4 shows the result of DCBA-DEA. To note, the number of simulation replications is set as 2000, this is because based on the analysis of Kao and Liu (2009), they indicate that 2000 simulation replications are sufficient to produce a result close to the true distribution.

From Table 4, we can find that in the method of using average data, banks 4, 11, 13, 14, and 23 are efficient, while in DCBA-DEA, we have missed the banks 13 and 23 for the lower bound of the efficiency scores are less than one. For those with the same upper efficiency bound (e.g., banks 3, 19, and 25), we are still unable to rank them based on their lower efficiency bound, we should know the distribution of their efficiency scores in the statistical propose. If two banks have almost the same efficiency ranges and the means, we should check the standard deviations and rank the one with smaller standard deviation as a better performance. The reason is just as we have discussed in the numerical example, the smaller standard deviation show more stability of the result.

Meanwhile, the numbers from Table 4 indicate that the average difference between the mean efficiency of each bank calculated from the simulations (column 7) and the corresponding efficiency calculated from the average data (column 2) is 0.019, which is approximately 2% of the average efficiency of the 25 banks. A Wilcoxon signed-ranks test shows that the efficiency scores obtained from two approaches are significant different at 0.05 level. While from Table 4, it can be observed that the rankings of the 25 banks are not so different although their efficiency scores are from different approaches. In fact, there are 12 banks whose ranks from 3 three approaches are exactly the same; in addition, there are only three banks (19, 22, and 25) with larger difference.

To note that from the perspective of efficiency ranges, our result is a little different from Kao and Liu (2009)'s, e.g., for bank 14, their efficiency range for bank 14 is [0.8429, 1.000] with mean efficiency of 0.9996,

Table 3. Data of 25 Taiwan commercial banks

Bank		No. of employees	Physical capital	Purchased funds	Demand deposits	Short-term loans	Medium-term loans
1	[L, U] Ave.	[6092, 6211] 6128	[21.5, 25.1] 22.5	[557.5, 754.8] 646.1	[107.4, 142.8] 129	[368.1, 469.9] 426.3	[196.4, 266.9] 234.7
2	[L, U] Ave.	[6481, 6782] 6630	[24.7, 27.6] 26.5	[681.1, 855.5] 763	[118.2, 177.9] 146.4	[407.8, 510.8] 473.8	[225.8, 293.9] 262.7
3	[L, U] Ave.	[6138, 6291] 6209	[20.5, 21.6] 21.3	[658.5, 795.9] 713	[122.8, 179.6] 151.9	[405.5, 542.6] 468.1	[190.8, 279.5] 233.3
4	[L, U] Ave.	[3304, 3574] 3477	[10.1, 11.3] 10.9	[310.7, 416.9] 351.4	[65.8, 109.8] 85.2	[257.6, 366.3] 309.1	[91.1, 130.8] 114
5	[L, U] Ave.	[1988, 2162] 2074	[6.0, 7.3] 6.6	[191.8, 249.3] 220.6	[21.0, 24.0] 22.4	[104.5, 140.9] 119.8	[39.9, 97.9] 64.6
6	[L, U] Ave.	[2005, 2176] 2117	[3.5, 4.9] 4	[192.2, 237.6] 218.2	[23.0, 31.5] 27.3	[128.5, 149.8] 138.6	[50.4, 63.2] 55.6
7	[L, U] Ave.	[1230, 1384] 1320	[3.2, 3.4] 3.3	[108.0, 166.8] 124.7	[5.9, 6.6] 6.2	[47.4, 60.8] 54.3	[28.1, 37.9] 32.2
8	[L, U] Ave.	[931, 1310] 1097	[3.0, 3.4] 3.2	[74.6, 84.1] 79.1	[4.0, 5.0] 4.4	[28.4, 36.7] 33.1	[21.4, 23.4] 22.4
9	[L, U] Ave.	[1994, 2300] 2144	[3.1, 4.1] 3.6	[162.7, 188.7] 174.9	[16.5, 24.1] 20.1	[95.6, 120.7] 107.1	[32.5, 40.0] 36
10	[L, U] Ave.	[3854, 5704] 4751	[16.9, 26.3] 21.3	[332.2, 477.5] 397.9	[24.5, 35.6] 30.4	[234.1, 321.9] 279	[90.8, 182.3] 133.5
11	[L, U] Ave.	[2194, 2398] 2294	[5.6, 6.0] 5.8	[285.9, 346.5] 316.9	[33.0, 91.9] 58.4	[186.1, 212.3] 200.6	[115.9, 148.4] 138.7
12	[L, U] Ave.	[930, 1446] 1201	[2.5, 5.4] 4.6	[104.5, 139.1] 123.2	[12.5, 17.5] 14.9	[84.0, 111.6] 99.6	[25.3, 39.1] 33
13	[L, U] Ave.	[977, 2000] 1323	[0.9, 2.1] 1.4	[106.0, 163.9] 139.2	[7.3, 13.8] 10.7	[93.9, 128.3] 116.4	[19.0, 30.8] 23.7
14	[L, U] Ave.	[1050, 1507] 1240	[0.6, 1.3] 1	[105.4, 183.1] 154.6	[6.1, 10.4] 7.4	[78.1, 139.5] 118.2	[21.7, 27.3] 24.7
15	[L, U] Ave.	[4905, 5438] 5194	[13.4, 15.7] 14.6	[539.1, 655.4] 592.2	[84.1, 104.2] 96	[323.8, 392.1] 351.9	[219.5, 271.3] 243.5
16	[L, U] Ave.	[984, 1980] 1568	[3.1, 4.5] 4.1	[110.5, 160.3] 142.8	[7.6, 11.0] 8.7	[96.9, 112.7] 105.3	[12.6, 25.5] 21.2
17	[L, U] Ave.	[1350, 1811] 1614	[1.6, 3.1] 2.4	[111.3, 150.9] 139.1	[8.9, 12.1] 10.7	[76.0, 111.8] 95.4	[20.5, 43.1] 32.4
18	[L, U] Ave.	[1015, 1869] 1508	[2.5, 4.5] 3.8	[120.1, 194.6] 155.2	[14.5, 22.7] 18.2	[71.9, 78.2] 74.4	[35.1, 87.4] 59.6
19	[L, U] Ave.	[856, 1458] 1195	[2.5, 4.7] 3.7	[121.5, 197.5] 165.9	[13.3, 20.6] 16	[89.8, 116.2] 105.6	[28.4, 59.0] 45.7
20	[L, U] Ave.	[1350, 1750] 1540	[3.1, 4.6] 3.5	[120.8, 209.7] 171.8	[10.5, 17.8] 15.1	[75.3, 97.9] 85.7	[41.8, 61.7] 53.6
21	[L, U] Ave.	[819, 1055] 981	[1.3, 1.7] 1.5	[84.2, 114.0] 105.2	[6.4, 8.4] 7.5	[59.3, 73.0] 67.7	[23.5, 36.4] 30.8
22	[L, U] Ave.	[1538, 3278] 2409	[3.1, 9.2] 5.2	[121.5, 219.7] 168	[8.1, 15.3] 11.3	[111.4, 153.2] 140.2	[13.7, 40.6] 26.9
23	[L, U] Ave.	[1035, 1484] 1230	[1.2, 2.3] 1.8	[85.5, 115.6] 104.1	[5.9, 12.1] 8.5	[64.1, 90.0] 79.7	[28.4, 46.0] 36.1
24	[L, U] Ave.	[1004, 1808] 1421	[2.7, 3.4] 3.1	[106.7, 151.5] 129.1	[5.9, 11.5] 8.2	[67.2, 88.7] 78.8	[35.2, 53.2] 45.2
25	[L, U] Ave.	[887, 1441] 1261	[0.5, 2.4] 2	[94.3, 161.1] 140.4	[7.1, 9.9] 8.3	[74.4, 100.9] 91.6	[22.5, 50.8] 39.4

while in DCBA-DEA, bank 14 is always efficient. To note that the mean efficiency of bank 14 of Kao and Liu

(2009) is 0.9996 while the lower bound is 0.8429, this indicates that their result is much wider than ours. Such

Table 4. Result of efficiency scores from different approaches

Bank	Average data		Kao and Liu (2009)			DCBA-DEA		
	efficiency	Rank	(mean, SD)	Efficiency range	Rank	(mean, SD)	Efficiency range	Rank
1	0.9411	8	(0.9255, 0.0483)	[0.7825, 1.000]	11	(0.9188, 0.0207)	[0.864, 0.9728]	11
2	0.8969	14	(0.9022, 0.0490)	[0.7634, 1.000]	14	(0.9003, 0.0119)	[0.8686, 0.9405]	14
3	0.9737	6	(0.9404, 0.0477)	[0.7860, 1.000]	9	(0.9768, 0.0091)	[0.9495, 1]	6
4	1	1	(1.000, 0.000)	[1.000, 1.000]	1	(1.0000, 0.0000)	[1, 1]	1
5	0.7444	21	(0.7829, 0.0721)	[0.6125, 1.000]	21	(0.7765, 0.0189)	[0.7201, 0.828]	21
6	0.8217	20	(0.8322, 0.0331)	[0.7123, 0.9790]	20	(0.8293, 0.0081)	[0.8043, 0.8507]	20
7	0.6291	25	(0.5885, 0.0429)	[0.4866, 0.8237]	25	(0.5842, 0.0239)	[0.5258, 0.6584]	25
8	0.6534	24	(0.6638, 0.0292)	[0.5575, 0.7805]	24	(0.6477, 0.0046)	[0.6362, 0.6637]	24
9	0.7404	23	(0.7430, 0.0351)	[0.6213, 0.9853]	23	(0.7416, 0.0078)	[0.717, 0.7667]	23
10	0.9082	11	(0.8970, 0.0660)	[0.7159, 1.000]	15	(0.8966, 0.0167)	[0.8508, 0.9422]	15
11	1	1	(1.000, 0.000)	[1.000, 1.000]	1	(1.0000, 0.0000)	[1, 1]	1
12	0.9286	10	(0.9404, 0.0534)	[0.7009, 1.000]	10	(0.9245, 0.0122)	[0.8913, 0.9635]	10
13	1	1	(0.9838, 0.0355)	[0.7641, 1.000]	5	(1.0000, 0.0001)	[0.9979, 1]	4
14	1	1	(0.9996, 0.0054)	[0.8429, 1.000]	3	(1.0000, 0.0000)	[1, 1]	1
15	0.9394	9	(0.9677, 0.0352)	[0.8125, 1.000]	7	(0.9422, 0.0087)	[0.9154, 0.9684]	9
16	0.8423	19	(0.8906, 0.0575)	[0.6748, 1.000]	16	(0.8831, 0.0145)	[0.8372, 0.9309]	17
17	0.8439	18	(0.8712, 0.0602)	[0.6863, 1.000]	18	(0.8731, 0.0173)	[0.8194, 0.9158]	18
18	0.8787	15	(0.9095, 0.0933)	[0.6279, 1.000]	13	(0.8896, 0.0313)	[0.7929, 0.9786]	16
19	0.8665	16	(0.9611, 0.0477)	[0.7988, 1.000]	8	(0.9740, 0.0157)	[0.9236, 1]	7
20	0.7431	22	(0.7786, 0.0656)	[0.6001, 1.000]	22	(0.7613, 0.0161)	[0.7142, 0.8052]	22
21	0.904	12	(0.9196, 0.0453)	[0.7614, 1.000]	12	(0.9181, 0.0130)	[0.8776, 0.9633]	12
22	0.9466	7	(0.8798, 0.0807)	[0.6412, 1.000]	17	(0.9066, 0.0432)	[0.8117, 1]	13
23	1	1	(0.9943, 0.0175)	[0.8282, 1.000]	4	(1.0000, 0.0005)	[0.9891, 1]	5
24	0.8653	17	(0.8599, 0.0546)	[0.7096, 1.000]	19	(0.8505, 0.0141)	[0.815, 0.8998]	19
25	0.8975	13	(0.9816, 0.0387)	[0.7556, 1.000]	6	(0.9489, 0.0386)	[0.877, 1]	8

DCBA-DEA: Data Collection Budget Allocation Data Envelopment Analysis, SD: standard deviation.

cases also appear in the efficiency measurement of other banks. This is because in our method, the efficiencies can be better estimated by collecting more relevant data, through an intelligent data collection plan, and DCBA-DEA helps to materialize this. Compared with Kao and Liu (2009), the data collection plan and the improvement of the reducing standard deviation and narrowing efficiency ranges are shown in Table 5.

From Table 5 it shows that the comparisons of statistic information with ranges and standard deviation from the method of Kao and Liu (2009) and DCBA-DEA. However, the efficiency scores from the latter method are more accurate than the former. This is reflected in the improvement in efficiency ranges and standard deviations have been reduced greatly. The 25 banks experience about the average reduction of 65.6% in efficiency ranges. Take bank 8 as an example, its efficiency interval is [0.5575, 0.7805] (range = 0.2230) from the approach of Kao and Liu (2009), and it be-

comes [0.6362, 0.6637] (range = 0.0275), which represents and improvement of about 84%. This implies that the true mean values of the efficiency scores are being estimated more accurately. Meanwhile, the standard deviation, as a significant indicator of simulation stability, has been greatly reduced by 66.2% in average. As we know from statistics, a smaller standard deviation can predict the efficiency scores into narrower ranges. Here for instance, the bank 8, its standard deviation has been reduced from 0.0292 to 0.0046, this means an improvement of standard deviation reduction of 87.67% by applying DCBA-DEA. In addition, the value of MSE, this is as a measurement of accuracy level in efficiency scores, appears at the third digit after the decimal point for all the banks.

For the conventional model that uses the average data to estimate efficiency, and each efficiency score is a discrete value. While for the stochastic models, apart from the mean efficiency scores, the standard deviation,

Table 5. Data collection design and the improvement of DCBA-DEA in ranges and standard deviation

Bank	DCBA-DEA data collection design and MSE							Improvement of accuracy	
	n1	n2	n3	n4	n5	n6	MSE	Ranges reduction ($\frac{*R1 - R2}{R1} \times 100\%$)	Standard reduction ($\frac{**S1 - S2}{S1} \times 100\%$)
	1	47	39	66	1	1	46	0.000745	57.14
2	27	8	24	82	15	44	5.90E-05	75.71	69.61
3	10	20	9	153	7	1	5.27E-05	80.92	76.40
4	26	48	8	106	9	3	4.37E-26	0.00	0.00
5	8	14	39	36	4	99	0.000738	73.79	72.15
6	18	29	86	24	21	22	6.55E-05	75.53	82.60
7	26	15	7	7	34	111	0.001137	44.29	60.66
8	27	35	53	19	18	48	2.53E-05	84.25	87.67
9	18	84	32	1	56	9	3.49E-05	77.78	86.35
10	7	11	61	11	47	63	0.000187	74.70	67.83
11	70	46	46	26	3	9	1.64E-27	0.00	0.00
12	80	1	43	1	74	1	8.71E-05	77.15	75.86
13	11	11	85	16	64	13	1.94E-18	99.72	99.11
14	18	44	56	45	24	13	7.89E-26	100.00	100.00
15	11	4	101	9	29	46	2.92E-05	75.28	71.73
16	60	1	42	5	75	17	0.001451	74.78	71.19
17	17	31	29	21	44	58	0.000746	71.26	69.27
18	19	24	93	3	6	55	0.000502	66.45	50.09
19	112	15	8	11	43	11	0.000168	67.09	62.03
20	18	17	69	5	42	49	0.000246	75.46	77.24
21	34	39	36	1	55	35	0.000166	71.30	64.08
22	47	7	62	68	1	15	0.001869	46.47	47.52
23	3	51	38	69	26	13	3.40E-23	97.14	93.66
24	1	1	55	1	90	52	0.000185	74.18	70.80
25	33	1	79	2	45	40	0.003060	0.26	49.67

DCBA-DEA: Data Collection Budget Allocation Data Envelopment Analysis, MSE: mean square error.

* R1: range of efficiency from Kao and Liu (2009), R2: range of efficiency from DCBA-DEA.

** S1: standard deviation from Kao and Liu (2009), S2: standard deviation from DCBA-DEA.

the efficiency ranges (minimum and maximum values) can be obtained from the results as well. What is more important, the approach of DCBA-DEA which can make the efficiency scores have a better estimation by collecting more relevant data intelligently, this is because that the data collection designs can direct the Monte Carlo simulation more sufficiently when compared with the normal simulation approach (e.g., Kao and Liu, 2009). When it comes to the reality, as user want to do future decision-making base on the estimation of the existing information (e.g., efficiency). A smaller standard deviation will give user a better prediction of company operations, this is due to the inputs (e.g.,

cost) can be limited in a smaller ranges to avoid the waste of cost. Then it will help the company to save cost and improve business efficiency. In addition, the results from DCBA are useful for managers to know how to effectively allocate the budget for data collection in order to maximize the accuracy of efficiency, then they can conveniently judge the performance optimistically (maximum values), moderately (mean values), or conservatively (minimum values) based on their experience, expertise, and judgment. Moreover, the data collection design, i.e., in Table 5, can express the randomness of the stochastic variables, the larger of the number indicate that this variables are more stochastic than the other,

and hence it need more simulation runs to estimate a value which is more close to its true value. Further, these data collection plans even can direct the real data collection process if more information (e.g., the objective, the total data collection plan, budget, and time) can be added.

5. CONCLUDING REMARKS

This research developed a method known as the DCBA-DEA for measuring efficiency for DMUs in a more practically feasible way. The proposed method starts by measuring the efficiency scores, after which it improves the accuracy of the efficiency through data collection. The effort in data collection is allocated intelligently using the simulation optimization technique.

The proposed method was designed to tackle the limitations of the conventional efficiency measurement modulus operandi. To that end, the first salient point is that efficiency scores are obtained in a more confident manner, as to which even the decision makers are uncertain about the data, the efficiency scores can still be estimated accurately and performance analysis can be conducted smoothly. The second point to consider is that the proposed method, when collecting the data, one will be able to allocate the effort intelligently, eliminating wastes of resources and not necessities in collecting data which are not necessary that might otherwise result in misleading efficiencies.

On the limitation, we admit that the experiment in Sections 4 and 5 was not a full-scale application, but rather a partial-scale study utilized just to illustrate the proposed method. The application has been simplified and various considerations need to be taken into account for a fully justifiable DCBA-DEA application. Nevertheless, it is hoped that, this paper could be the breaking point for a new direction in efficiency measurement research, i.e., to employ 'thinking out of the box' approach for more innovative and creative ideas.

ACKNOWLEDGMENTS

The authors acknowledge the grant support by RCMO USM (1001/PMGT/816224) and ERGS (203/PMGT/6730 127). The first author acknowledges the funding support from School of Management and fellowship support from the Institute of Postgraduate Studies, Universiti Sains Malaysia.

REFERENCES

Banker, R. D., Kotarac, K., and Neralic, L. (2013), Sensitivity and stability in stochastic data envelopment analysis, *Journal of the Operational Research So-*

ciety, **65**(4), 1-14.

Charnes, A., Cooper, W. W., and Rhodes, E. (1978), Measuring the efficiency of decision making units, *European Journal of Operational Research*, **2**(6), 429-444.

Chen, C. H. and Lee, L. H. (2010), *Stochastic Simulation Optimization: An Optimal Computing Budget Allocation*, World Scientific Publishing Co., Singapore.

Cooper, W. W. (2013), data envelopment analysis. In: Gass, S. I. and Fu, M. C. (eds.), *Encyclopedia of Operations Research and Management Science*, Springer, Heidelberg, 349-358.

Cooper, W. W., Huang, Z., and Li, S. X. (1996), Satisficing DEA models under chance constraints, *Annals of Operations Research*, **66**(4), 279-295.

Cooper, W. W., Seiford, L. M., and Tone, K. (2000), *Data Envelopment Analysis: A Comprehensive Text with Models, Applications, References, and DEA-Solver Software*, Kluwer Academic, Boston, MA.

Despotis, D. K. and Smirlis, Y. G. (2002), Data envelopment analysis with imprecise data, *European Journal of Operational Research*, **140**(1), 24-36.

Dyson, R. G. and Shale, E. A. (2010), Data envelopment analysis, operational research and uncertainty, *Journal of the Operational Research Society*, **61**(1), 25-34.

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

Kao, C. and Liu, S. T. (2009), Stochastic data envelopment analysis in measuring the efficiency of Taiwan commercial banks, *European Journal of Operational Research*, **196**(1), 312-322.

Lee, L. H., Chew, E. P., Frazier, P. I., Jia, Q. S., and Chen, C. H. (2013), Advances in simulation optimization and its applications, *IIE Transactions*, **45**(7), 683-684.

Ogawa, M. and Ishii, H. (2002), The total ranking method from multi-categorized voting data based on analytic hierarchy process, *Industrial Engineering and Management Systems*, **1**(1), 93-98.

Okudan, G. E. and Lin, C. Y. (2009), An exploration on the use of data envelopment analysis for product line selection, *Industrial Engineering and Management Systems*, **8**(1), 47-53.

Olesen, O. B. (2006), Comparing and combining two approaches for chance constrained DEA, *Journal of Productivity Analysis*, **26**(2), 103-119.

Olesen, O. B. and Petersen, N. C. (1995), Chance constrained efficiency evaluation, *Management Science*, **41**(3), 442-457.

Panta, M., Smirlis, Y., and Sfakianakis, M. (2013), Assessing bids of Greek public organizations service

- providers using data envelopment analysis, *Operational Research*, **13**(2), 251-269.
- Seiford, L. M. (1996), Data envelopment analysis: the evolution of the state of the art (1978–1995), *Journal of Productivity Analysis*, **7**(2/3), 99-137.
- Simar, L. and Wilson, P. W. (1998), Sensitivity analysis of efficiency scores: how to bootstrap in nonparametric frontier models, *Management Science*, **44**(1), 49-61.
- Thanassoulis, E. (2001), *Introduction to the Theory and Application of Data Envelopment Analysis: A Foundation Text with Integrated Software*, Kluwer Academic Publisher, Norwell, MA.
- Trinh, T. H., Kachitvichyanukul, V., and Luong, H. T. (2012), A tradeoff between customer efficiency and firm productivity in service delivery systems, *Industrial Engineering and Management Systems*, **11**(3), 224-232..
- Wong, W. P., Jaruphongsa, W., and Lee, L. H. (2008), Supply chain performance measurement system: a Monte Carlo DEA-based approach, *International Journal of Industrial and Systems Engineering*, **3**(2), 162-188.
- Wong, W. P., Jaruphongsa, W., and Lee, L. H. (2011), Budget allocation for effective data collection in predicting an accurate DEA efficiency score, *IEEE Transactions on Automatic Control*, **56**(6), 1235-1246.
- Zhu, J. (2003), Imprecise data envelopment analysis (IDEA): a review and improvement with an application, *European Journal of Operational Research*, **144**(3), 513-529.