Available online at http://scik.org

J. Math. Comput. Sci. 11 (2021), No. 1, 703-715

https://doi.org/10.28919/jmcs/5205

ISSN: 1927-5307

A NEW STEPWISE METHOD FOR SELECTION OF INPUT AND OUTPUT VARIABLES IN DATA ENVELOPMENT ANALYSIS

T. SUBRAMANYAM[1,*], RANADHEER DONTHI[2], V. SATISH KUMAR[3], S. AMALANATHAN[4], MADHUSUDHAN ZALKI[5]

[1]Department of Mathematics & Statistics, M.S. Ramaiah University of Applied Sciences, Bangalore 560054, India

[2]Department of Mathematics, St. MARTIN'S Engineering College, Telangana 500100, India

[3]Department of Computer Science, East Point College of Higher Education, Bangalore 560049, India

[4]Department of Business and Management, Christ (Deemed to be University), Bangalore 560029, India

[5]Department of Mathematics, REVA University, Bangalore 560064, India

Abstract: Data envelopment analysis (DEA) is one of the widely accepted optimization technique uses to measure the relative efficiency of organizational units where multiple inputs and outputs are present. The significance of DEA results depends on the variables selected for DEA modelling. One of the main challenges in data envelopment analysis modelling is of identify the significant input and output variables for DEA modelling. In this study, we propose an enhanced stepwise method to identify the significant and insignificant input and output variable by reducing the iterations process in stepwise method. The statistical significance of the input and output variables evaluated using the statistical methods: Least significance difference (LSD), and Welch's statistics. The proposed method applied to the Indian banking sector and the results have shown that the proposed model significantly identified the significant and

*Corresponding author

E-mail address: subramanyam.mt.mp@msruas.ac.in

Received November 14, 2020

insignificant input and output variables with least loss of information.

Keywords: stepwise method; commercial banks; data envelopment analysis; dimensionality; Welch's statistic.

2010 AMS Subject Classification: 90C08, 90B50.

1. INTRODUCTION

Data envelopment analysis is an optimization technique used to measure the relative efficiency of organizational units called decision-making units (DMUs), where multiple numbers of input and output variables are present during efficiency evaluation. The idea of technical and allocative efficiency was first originated by Farrell (1953) and subsequently, the fundamental models were developed by Charnes et.al (1978, 1984). These models create a frontier using the available input and output variables. The DMUs on the frontier line are called efficient (best practices) with efficiency score 1 and others are termed as inefficient with the efficiency score between 0 and 1. The major advantage of DEA is that it is data-driven and the relative weights of the variables need not be known a priori. The fundamental DEA models were applied in different fields like production management, banking, agriculture etc. to evaluate the relative efficiency of the branches and organizational units [3, 7, 10, 17, 18].

Modelling of DEA depends on researchers' perspective due to the lack of the standard procedures for the selection of input and output variables. The modelling of DEA never discusses how to identify the relevant input and output variables and assumes that the variables are known a priori. Since, the DEA analysis relies heavily on the selected input and output variables there are huge variations among the efficiency scores of DMUs from one researcher to another researcher. Due to the inclusion of more number of input and output variables, the dimensionality of the production possibility set will increase, and proportionally it leads to the poor discriminatory power of the DEA models. The general guidelines in DEA about the number of input and output variables is that the total input and output variables must be less than one third of the total DMUs [6, 13, 16]. Most of the researchers applied DEA models with the assumption that the input and output variables are known a priori. If the more number of input and output variables are present during

DEA modelling, some of the variables may not have significant impact on the efficiency scores of DMUs. Such variables don't have any role in DEA modelling but may decrease the discretionary power of DEA models. To avoid this situation, some of the researchers proposed variable selection methods based on statistical approaches. The rationale to develop these procedures is to identify the insignificant variables for improving the discriminating power of DEA. The usefulness of any DEA model will depend on the variables that are selected during the efficiency evaluation.

2. REVIEW OF LITERATURE

In DEA modelling, usually the input and output variables are highly correlated due to their interrelationships. Removing the variables using simple correlation and regression may not be possible. There are some studies discussed about the importance of correlation and regression methods, and Principle Component Analysis. In this approach the variables which are highly correlated with existing model variables are merely redundant and are omitted from further analysis [9, 13, 16]. Jenkins et.al, (2003) discussed a multivariate statistical approach to identify the variables that can be eliminated using least loss of information. JM Wagner (2007) proposed a stepwise method for identifying the insignificant variables based on the average changes when the variables are dropped from the data exploration. The inclusion of variables left to the researcher's discretion and this method unable to explain the proper cut off points for excluding the variables from the data exploration. The selection of variables for inclusion and elimination is purely based on judgmental method.

Nataraja N.R. et.al, (2011) outlined various variable selection techniques available in literature for DEA modeling. This study demonstrated various methods with statistical approaches like principle component analysis, regression-based tests, and the tests based on bootstrapping methods and proved that for the given data ECM approach works better for low correlated variables. Subramanyam T (2016) proposed a stepwise method to eliminate the insignificant variables with statistical significance using independent sample t-test. This approach tested the significance of the difference of means before and after elimination of the variables. Among all the insignificance

variables the variable with least loss of information is dropped from the analysis. This paper demonstrated the significance of the full and reduced model and the change of pure technical efficiency and scale efficiency with an empirical study using the Indian banking system.

Limleamthong, P et.al (2017) proposed a MIP-DEA model as a bi-level model to identify the insignificance variables. The outer model selects the metrics to be eliminated and the inner model calculates the efficiency scores. This model demonstrated for different case studies for food waste management technologies, electricity generation technologies and solvents for CO_2 capture. Li, Y. Et.al (2017) proposed a method to identify the proper input and output variables set for evaluation via Akaike's information criteria (AIC) rule. This method mainly focused on assessing the importance of subset of original variables rather than testing the marginal role of input and output variables one by one like other proposed models. Wilson, P. W. (2018) explains the dimension reduction and their use in efficiency evaluation for free-disposal models and DEA models. The simulation techniques applied to test the dimension reduction process and demonstrate that dimension reduction is advantageous in terms of reducing estimation error. Eskelinen, J. (2017) compared two different variable reduction procedures proposed by Jenkins and Anderson (2003) and Pastor, Ruiz, and Sirvent (2002) in an empirical retail bank context. The efficiencies obtained by using these methods were diverged. It is suggested that these techniques can be utilized to evaluate the banks from multiple perspectives.

The proposed models in literature discuss how to identify the insignificant variables and decrease the dimension of the DEA models. There is no study about the identification of significant variables and reduction of iteration process. This paper advances the work by proposing a stepwise method in two-way direction. i.e., the method identifies significant and insignificant variables in each of iteration to reduce the iteration process. Both significant and insignificant variables identified using the advanced statistical approaches like multiple comparison tests and Welch's statistics. The model suggests some simple rules to fix the variables for further analysis and strong statistical cut off for variable deletion from the data exploration. The backward elimination process proposed to delete the insignificant variables from the data exploration. Also, this proposed model is suitable

and understandable for all researchers using DEA with little mathematical knowledge.

3. BASIC CCR MODEL

Charnes et al., (1978) formulated a fractional programming problem to measure the efficiency of organizational units where multiple number of inputs produces multiple numbers of outputs. Assume that there are 'n' decision-making units producing 'O-outputs' using 'I-inputs' in a similar working environment. Where 'i' and 'r' assumes the values i=1, 2…, I and r=1, 2… O and 'j' assumes the values j=1, 2, ..., n. The basic CCR model can be represented as:

$$Z(CCR) = Min\left\{\lambda: \sum_{j=1}^{n} \lambda_j x_{ij} \le \lambda x_{i0};\ \sum_{j=1}^{n} \lambda_j u_{rj} \ge u_{r0};\ \lambda_j \ge 0,\ i = 1,2,\ldots,I;\ r = 1,2,\ldots,O\right\}$$

Stepwise Method: Backward Elimination Method:

The basic CCR model explained in section (3), utilized to run the proposed step-wise algorithm. While applying any statistical methods there are some basic assumptions. We considered the following assumptions for developing the proposed procedure as:

Assumptions:

1. The data for all input/output variables is available and are always greater than 'zero'.
2. Always there must be at least one input and output variable in the data exploration.
3. Only one input/output variable be removed at a time from the data exploration.
4. The input/output variable eliminated from the data exploration will not be included.
5. The efficiency scores follow normal distribution.

Under the necessary assumptions stated above, the following stepwise procedure proposed to select the significant variables and to remove the insignificant variables from the data exploration.

Step1: Run full model with all available input/output variables and store the efficiency scores in a set 'OTE'.

Step2: Drop all input and output variables one by one (if no variables fixed) and run the models. Store the efficiency scores in another set 'E_k'. Where 'k' takes the values from 1 to K=I+O. (k= 1, 2, ... , K)

Step3: Use Fisher's protected "Least Significant Difference (LSD)" method to test the significance difference between the means at 10% level of significance. The LSD uses the formula:

$$LSD(0.10) = t_{0.10} * s_d$$

Here $s_d = \sqrt{\frac{2s^2}{n}}$, n is number of observations, s^2 is the mean square error.

Step4: Retain all the input/output variables which are significant at 10% level of significance. No further significance test required for this variable(s).

Step5: Calculate the percentage of average change and variability change for full and reduced models respectively. Here, M and SD represents the mean and standard deviation respectively.

$$Mean\ change = \left|\frac{M_{Reduced\ Model(k)} - M_{Full\ Model}}{M_{Full\ Model}}\right| * 100$$

$$Change\ of\ Variability\ = \left|\frac{SD_{Reduced\ Model(k)} - SD_{Full\ Model}}{SD_{Full\ Model}}\right| * 100$$

Step6: The variable which is statistically insignificant with least loss of information (i.e., least average and variability change) will leave from the data exploration.

Step7: Repeat step-1 to step-7 until all variables are statistically significant in the data exploration. The final variables in DEA modelling are purely based on researcher's discretion.

Step8: Under the assumption of normality, use Welch's t-statistic to test the significance difference between the average scores of full and reduced model.

$$W = \left|\frac{M_{Reduced\ Model} - M_{Full\ Model}}{SE(M_{Reduced\ Model} - M_{Full\ Model})}\right| \sim t_{(N-2,0.10)}$$

Step9: **Final Reduced Model:**

(a) If there is no statistical significant difference between the averages of full and reduced models, the reduced model is an appropriate model for further analysis.

(b) If there is a significant difference between the averages of full and reduced models, add the dropped variables from steps K to 1, until there is no significant difference between full and reduced model. (This step left to users' discretion).

4. INDIAN BANKING SYSTEM

Indian banking system is one of the strong and stable industry comparing to any other countries' banking system. This sector plays a major role in the growth of Indian economy. The 'Reserve Bank of India (RBI)' is the monitoring authority of all banks in India and regulates the banking business of all the banks according to the needs of the Indian economy. Due to the globalization, more number of private and foreign sector banks started working in India [8, 14, 15]. In India, the banking management system is broadly classified into three different categories based on the ownerships as Public, Private, and Foreign Sector banks. Gauging the efficiency of any commercial bank is important to the investors, policymakers and for a layman to know whether the banks are working in efficient environment or not due to the profound competition in banking business.

The DEA models were applied by number of researchers in evaluating the efficiency of banks and bank branches. Most of the studies evaluated the efficiency by assuming that the input and output variables are known a priori [8, 14, 17]. There is no general agreement on the modelling of DEA due to the availability of more number of input and output variables in banking business [9, 15, 16]. To identify a parsimonious model, number of researchers proposed different methods for reducing the input and output variables. The data is collected from the RBI Bulletins for the financial year, 2018-19 for all the public and private sector banks working in India.

5. INPUT AND OUTPUT VARIABLES

The study assumes production approach with the variables (i) Total number of employees working in each bank (ii) The fixed assets of each commercial bank and (iii) Total expenditure of banks as input variables and (i) Deposits, (ii) Investments (iii) Advances (iv) Interest Income and (v) Other income as output variables.

6. EMPIRICAL ANALYSIS

The present study conducted using the data of 42 Indian commercial banks comprising of 20 public

and 22 private sector banks for the financial year 2018-19. The correlation matrix among the selected input and output variables is represented in table (1). The table reveals that there is a high correlation among the input and output variables in DEA.

Table (1): Correlation Matrix

	No. of Employees	Fixed Assets	Total Exp.	Deposits	Investments	Advances	Interest income	Other income
No. of Employees	1							
Fixed Assets	0.9281	1						
Total Expenditure	0.9869	0.9496	1					
Deposits	0.9852	0.9585	0.9965	1				
Investments	0.9804	0.9606	0.9955	0.9957	1			
Advances	0.9883	0.9350	0.9976	0.9943	0.9894	1		
Interest income	0.9885	0.9326	0.9979	0.9932	0.9903	0.9987	1	
Other income	0.9741	0.8773	0.9719	0.9574	0.9547	0.9774	0.9789	1

Step-wise method: The basic CCR (1978) under the constant returns to scale utilized to determine the efficiencies of selected banks to run the proposed backward elimination method. In iteration-1, the method started with full model consisting of 3-input and 5-output variables. The average overall efficiency is 0.9515 with 18 efficient banks.

Table (2): Summary of Iteration-I

	Average Efficiency	Mean Difference	Average Change (%)	Change of Variability (%)	Remarks
Overall Efficiency (3I,5O)	0.9515				
Variables Dropped					
Employees	0.9379	0.0136	-1.43	-12.76	
Fixed assets	0.9451	0.0065	-0.68	0.17	Dropped
Total expenditure	0.8103	**0.1412***	-14.84	-138.62	Sig.
Deposits	0.9365	0.0150	-1.58	-3.17	
Investments	0.9403	0.0112	-1.18	-6.19	
Advances	0.9492	0.0023	-0.24	-0.38	
Interest income	0.9384	0.0131	-1.38	-25.81	
Other income	0.9402	0.0113	-1.19	1.02	

The stepwise method started with dropping one variable at a time. The efficiencies were calculated and the Fisher's LSD method reveals that only 'total expenditure' having significant impact. From

other insignificant variables the variable 'fixed assets' has less impact on the average change and variability change. The variable 'total expenditure' fixed for further analysis and 'fixed assets' dropped from the data exploration.

Table (3): Summary of Iteration-2:

	Average Efficiency	Mean difference	Average change (%)	Change of variability (%)	Remarks
Overall Efficiency (3I,5O)	0.9515				
Variables Dropped					
Employees	0.9213	**0.0302***	-3.17	-9.94	Sig.
Deposits	0.9339	0.0176	-1.85	-6.60	
Investments	0.9353	0.0163	-1.71	-5.86	
Advances	0.9419	0.0096	-1.01	0.75	Dropped
Interest income	0.9262	0.0253	-2.66	-28.55	
Other income	0.9319	0.0197	-2.07	2.84	

In iteration-2, the stepwise method started by dropping one variable at a time. The only one variable 'Employees' seems to be statistically significant and this variable fixed for further analysis. Among the insignificant variables, the 'advances' has less impact comparing to all other variables. Therefore, the output variable 'advances' dropped from the data exploration. Next iteration started with 2-inputs and 4-output variables.

Table (4): Summary of Iteration-3:

	Average Efficiency	Mean difference	Average change (%)	Change of variability (%)	Remarks
Overall Efficiency(3I,5O)	0.9515				
Variables Dropped					
Deposits	0.9173	**0.0343***	-3.60	-8.25	Sig.
Investments	0.9331	0.0184	-1.93	-5.53	Dropped
Interest income	0.9094	**0.0422***	-4.43	-33.55	Sig.
Other income	0.9260	**0.0255***	-2.68	3.29	Sig.

In this stage, CCR model started with two fixed input variables namely, number of employees and total expenditure, and tested for the significance of four output variables. All output variables are

statistically significant except the variable 'Investments'. The average change and the variability change is also less comparing to other variables. Therefore, the output variable 'Investments' dropped from the data exploration.

6. STATISTICAL SIGNIFICANCE

To test whether there is any statistical significance difference between the full and reduced models with the null hypothesis $H_0: \mu_1 = \mu_2$, the Welch's statistics applied and shown that there is no statistical significance ($p > 0.10$). The final input and output variables will become as:

Table (5): Final Input and Output variables

Input Variables	Output Variables
1. Total employees	1. Deposits
2. Total expenditure of banks	2. Interest income
	3. Other income

All the above input and output variables are statistically significant at 10% level of significance. It is worth to assume 10% since we are dealing with highly correlated input and output variables which are involved in data envelopment analysis exploration.

7. SUMMARY AND CONCLUSIONS

The efficiency scores of DEA will be more accurate when the relevant significant input and output variables are considered for DEA modelling. Due to the availability of more number of input and output variables, interrelationships among the variables, it is difficult to identify the irrelevant variables which are having less impact on the efficiency scores of DEA. Number of studies proposed different methods on identifying the insignificant variables and dimensionality reduction. The present study advances the work by proposing a stepwise method in two-way direction. i.e., the method identifies significant and insignificant variables in each of iteration to reduce the iteration process using advanced statistical approaches like multiple comparison tests and Welch's statistics. The multiple comparison tests, namely, Fisher's Protected Least Significant Difference

(LSD) method used to identify the significant variables. The average change and average variability are used to identify the insignificant variables to leave from the data exploration. The proposed stepwise method applied for 42 commercial banks with 3-input variables and 5-output variables. During the data exploration using stepwise method, the variables fixed assets, advances and investments become insignificant and are eliminated from the data exploration. We observed that there is no statistical significance difference between the means of full and reduced models. It means, statistically, both the models are providing almost same information on the efficiencies of banks.

CONFLICT OF INTERESTS

The authors declare that there is no conflict of interests.

REFERENCES

[1] R. D. Banker, A. Charnes, W. W. Cooper, Some models for estimating technical and scale inefficiencies in data envelopment analysis, Manage. Sci. 30 (1984), 1078-1092.

[2] A. Charnes, W. W. Cooper, E. Rhodes, Measuring the efficiency of decision making units, Eur. J. Oper. Res. 2 (1978), 429-444.

[3] A. Colbert, R. R. Levary, M. C. Shaner, Determining the relative efficiency of MBA programs using DEA, Eur. J. Oper. Res. 125(3) (2000), 656-669.

[4] J. Eskelinen, Comparison of variable selection techniques for data envelopment analysis in a retail bank, Eur. J. Oper. Res. 259 (2017), 778-788.

[5] M. J. Farrell, The Measurement of Productive Efficiency, J. R. Stat. Soc. Ser. A (General). 120 (1957), 253-290.

[6] L. Friedman, Z. Sinuany-Stern, Combining ranking scales and selecting variables in the DEA context: The case of industrial branches, Comput. Oper. Res. 25 (1998), 781-791.

[7] L. Jenkins, M. Anderson, A multivariate statistical approach to reducing the number of variables in data envelopment analysis, Eur. J. Oper. Res. 147 (2003), 51-61.

[8] S. Kumar, R. Gulati, An examination of technical, pure technical, and scale efficiencies in Indian public sector banks using data envelopment analysis, Eurasian J. Bus. Econ. 1 (2008), 33-69.

[9] L. M. Seiford, J. Zhu, Modeling undesirable Factors in efficiency evaluation, Eur. J. Oper. Res. 142 (2004), 16-20.

[10] A. Y. Lewin, R. C. Morey, T. J. Cook, Evaluating the administrative efficiency of courts, Omega, 10 (1982), 401-411.

[11] Y. Li, X. Shi, M. Yang, L. Liang, Variable selection in data envelopment analysis via Akaike's information criteria, Ann. Oper. Res. 253 (2017), 453-476.

[12] P. Limleamthong, G. Guillén-Gosálbez, Mixed-integer programming approach for dimensionality reduction in data envelopment analysis: application to the sustainability assessment of technologies and solvents, Ind. Eng. Chem. Res. 57 (2018), 9866-9878.

[13] N. R. Nataraja, A. L. Johnson, Guidelines for using variable selection techniques in data envelopment analysis, Eur. J. Oper. Res. 215 (2011), 662-669.

[14] C. S. Reddy, T. Subramanyam, Data Envelopment Analysis Models to Measure Risk Efficiency: Indian Commercial Banks, IUP J. Appl. Econ. 10 (2011), 40-69.

[15] T. Subramanyam, C. S. Reddy, Measuring the risk efficiency in Indian commercial banking–a DEA approach, East-West J. Econ. Bus. 11(2008), 76-105.

[16] T. Subramanyam, Selection of input-output variables in data envelopment analysis-Indian commercial banks, Int. J. Computer Math. Sci. 5 (2016), 2347-8527.

[17] T. Subramanyam, B. Venkateswarlu, G. Mythili, R. Donthi, V. S. Kumar, Assessing the Environmental Efficiency of Indian Commercial Banks - An application of Data Envelopment Analysis, Adv. Math.: Sci. J. 9 (2020), 8037-8045.

[18] B. Venkateswarlu, T. Subramanyam, Efficiency evaluation of total manufacturing sectors of India – DEA approach, Glob. J. Pure Appl. Math. 11 (2015), 3145–3155.

[19] J. M. Wagner, D. G. Shimshak, Stepwise selection of variables in data envelopment analysis: Procedures and managerial perspectives, Eur. J. Oper. Res. 180 (2007), 57-67.

[20] P. W. Wilson, Dimension reduction in nonparametric models of production, Eur. J. Oper. Res. 267 (2018), 349-367.