Methodology Note

Measuring Hospital Performance through Data Envelopment Analysis: Understanding Basic Concepts to Help Novice Researchers

P.R. Sodani and G.M.K. Madnani

Attempting to assess the performance of organisations that are health service providers is not only a difficult task, but also requires specialised evaluation techniques because of the non-parametric measurable characteristics of their product. Data envelopment analysis (DEA) is a unique technique used for studying the comparative operating efficiency of an organisation in terms of scale and technical efficiencies that are beyond the scope of conventional measures of performance. The knowledge of these efficiencies reveal those 'black spots' that can then be targeted for improvement. Though being a very powerful methodology for evaluation of health service providers, DEA is not observed to have attracted the interest of research scholars and professionals working in the field of health care. Perhaps, it may be because of incomprehensible mathematical derivations that seldom provide insight as to how, why and what is actually happening in the shadows of such complicated manipulations. The purpose of this article is merely to unwind the minutely spun structure of the model and provide researchers a 'feel' for the hidden fibres of operating efficiencies that are at the very core of DEA model. By doing so we hope to help novice researchers in the field of health care, who are less accustomed with quantitative techniques, yet are more alive and knowledgeable about the concepts dealt in this methodology.

The Indian economy appears to have traversed from a rate of growth averaging 4.5 per cent to 8.5 per cent during the present decade. While celebrations over this remarkable transition from 'Hindu growth' to a new growth trajectory still continues, the evidence suggests that service and manufacture are the only two sectoral drivers that have given the present boost to the Indian economy. This paradigm shift to knowledge-based economy from an agrarian

Journal of Health Management, **10, 1 (2008): 129–142**
SAGE Publications Los Angeles • London • New Delhi • Singapore
DOI: 10.1177/097206340701000108

one brings larger responsibilities for these growth sectors; not only to sustain this rising trend, but also to evolve a framework that is well grounded in prudential principles of continual retaining quality with increased productivity. To achieve this, it is needed to strengthen, raise and evaluate performance to compete in the existing globalised competitive environment.

Of the different investigating techniques and approaches used as performance measures in service organisations, data envelopment analysis (DEA), combined with window analysis and the Malmquist productivity index, has received numerous applications in the last two decades with significant applications in education, hospitals, courts and banking institutions.

The present article lays emphasis on the basic concepts underlying the DEA approach in order to help our novice researchers understand the very crux of this evaluation methodology without going into complicated mathematical derivations with which they are less familiar.

What is Data Envelopment Analysis?

Based on the pioneering work of Farrell (1957), Charnes et al. (1978; the CCR model) and Banker et al. (1984; the BRR model) developed this non parametric mathematical programming model to its present operational level. This approach is used to evaluate relative performance of the members or decision-making units (DMUs) of a group that uses an identical set of inputs to produce a variety of identical outputs; but where the form of the production is not known or specified, as is the case with a parametric one.

The intrinsic value of this unique macro-analytical methodology lies in the fact that it reduces multiple inputs and outputs to single 'virtual' input and single 'virtual' output, regardless of different measurement units and requirement of standardisation. The relative efficiency of each DMU is then described by a simple ratio of virtual output to virtual input. The highest efficiency score is 1. The most efficient DMU among all the members in the reference group gets a score of 1, while those for which score is less than 1 are considered relatively inefficient.

Different Aspects of Efficiency

Two major aspects of overall efficiency with which every productive operation is concerned are technical efficiency and allocative efficiency. When a firm is able to generate maximum output from a given set of inputs, it is said to be technically efficient. On the other hand, when a firm is able to choose from different technically efficient sets of inputs, the one which costs the least it is said to be allocatively efficient. These two efficiencies are not mutually exclusive, nor are they mutually inclusive, yet every firm or organisation aspires to be both allocatively and technically efficient simultaneously to stay productively efficient overall. This riddle can be understood through simple principles of production economics.

A production process that uses two inputs (x and y) to produce a single product (Q) is described by a production function:

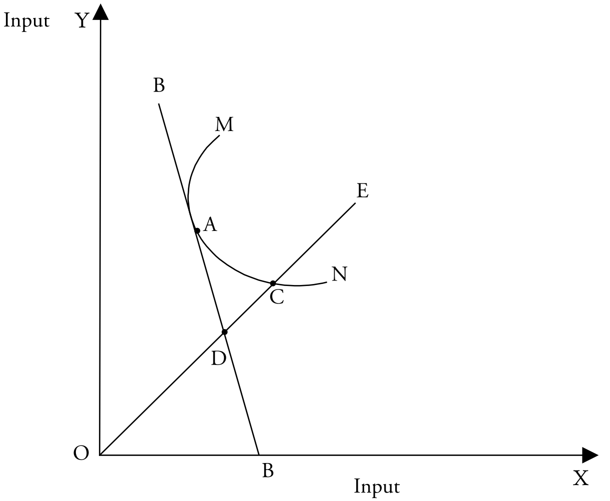
$$Q = \mathrm{f}\,(x, y),$$

and can also be diagrammatically expressed by an isoquant. An isoquant curve (or frontier) is the locus of different input combinations through which (on given technology) same level of output can be acquired.

Assume that a hypothetical DMU produces 100 units of product (Q) using two different inputs is represented by an isoquant *MN* of Figure 1. Every input combination on isoquant *MN* is considered technically efficient because any input combination of the *MN* isoquant can produce 100 units of Q; as such the DMU is said to be technically efficient when it is either on point *A* or point *C*. On the other hand, given market input prices and total affordable expenditure of DMU, *BB*' depicts its budget frontier. Any input combination beyond this frontier is not affordable by the DMU. Hence, tangency between isoquant and budget line is both technically and allocatively efficient point; which is point *A* in Figure 1. However, on point *E*, the DMU is both technically as well as allocatively inefficient, while on point *C* it is only allocatively inefficient.

On moving from *E* to *D*, the DMU's overall cost (or productive) efficiency will rise by (*OE-OD*)/*OE,* that is, say, the DMU's overall productive inefficiency will decrease by (*DE*/*OE*). This rise in the overall efficiency is due to gain acquired on two fronts: (*a*) that its technical efficiency increases by (*OE-OC*)/*OE* or the DMU's technical inefficiency decreases by (*CE*/*OE*); and (*b*) that its allocative efficiency increases by (*OC-OD*)/*OC* or the DMU's allocative inefficiency decreases by (*DC*/*OC*). Hence, the three points *A*, *C* and *E* (in Figure 1) describe three different situations of efficiency acquirement by the DMU: (*a*) on point *A* it is allocatively and technically efficient; (*b*) on point *C* it is only technically efficient, but not allocatively; and (*c*) on

Figure 1
Isoquant and Budget Frontier



point E it is allocatively inefficient in choosing an unaffordable input combination (= OD/OE) and also technically inefficient (= OC/OE). Both inefficiencies have led to excessive usage of two inputs (x and y) compared to when the DMU is on point A to produce 100 units of (Q).

Scale Efficiency

Other than the two efficiencies discussed previously, there is another one that exerts a direct impact on the overall efficiency of the DMU. With the increase in the scale of production, every DMU experiences economies or diseconomies of scale, which are integrated into technical efficiency. Hence, on raising the scale of operation beyond optimum scale of production (constant returns to scale, or CRTS), every DMU is likely to be affected by scale diseconomies or inefficiencies. Technical efficiency, therefore, shall not express the technological advantages in pure terms unless scale advantages are discounted from it. That is to say, the technical efficiency net of scale efficiency is the real or pure technical efficiency (PTE) of every DMU. This may be explained by taking an illustration of a hypothetical DMU that produces a single product by adding varying amounts of an input to fixed inputs.

Figure 2 describes the law of variable proportions and constant return frontiers through a continuous curve (2a) and kinked curve (2b). *OABC* depicts CRTS while *OPBV* and *SPBT* depict Variable Return to Scale (VRTS) frontiers. Point *E* is neither on the CRTS frontier nor on VRTS.

Figure 2 (a)
Variable and Constant Returns Frontiers

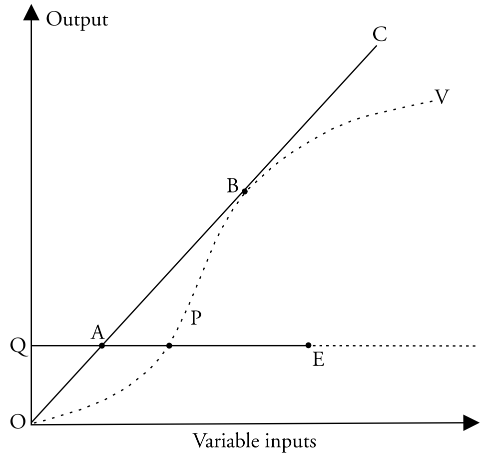
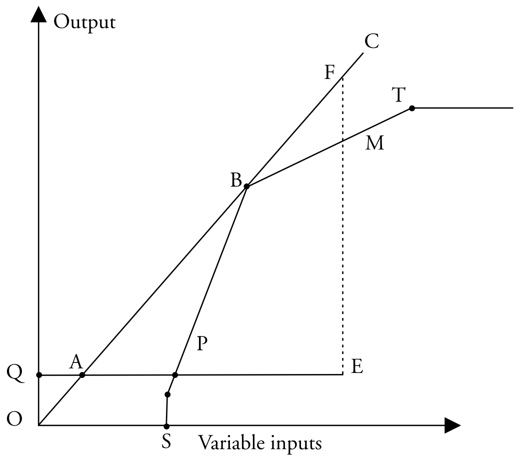


Figure 2 (b)



With reference to the CRTS frontier, point *E* is technically inefficient by distance *AE*; while with reference to the VRTS frontier this point (*E*) is also technically inefficient by distance *PE*. The difference between these two measures, *AP*, is attributable to scale inefficiency, which indicates that at *E* the DMU could produce its current level of output (say, *Q*) with lesser input in case it attains point *A*. Point *A* is on the optimum scale of production. The distance *PE* is, therefore, termed as pure technical inefficiency. Hence, overall technical inefficiency of the DMU is made up of two inefficiencies: scale inefficiency (*AP*) and pure technical inefficiency (*PE*).

Scale inefficiency arises because the DMU uses larger amount of input for output equivalent to the CRTS frontier, while pure technical inefficiency is due to failure to extract the maximum output from the present use of input by the DMU. Both inefficiencies describe unproductive use of resources.

Efficiencies can also be expressed in output maximisation or in output-oriented version. These concepts of efficiencies are based on input version. In the output maximisation version, distance *FE* is technical inefficiency of which *FM* is scale inefficiency; the remaining *ME* distance reflects pure technical inefficiency of the DMU.

The following discussion explains how performance efficiencies are brought into the envelopment analysis.

Data Envelopment Analysis Model

The approach behind this model is to use a linear programming procedure to investigate the performance of each DMU relative to performance of others of a group. Each member DMU is assigned a composite efficiency score on the basis of adopted input-output configurations. The DMU is then evaluated as relatively efficient (inefficient) if its calculated efficiency score is more (less) than the corresponding scores of the other units.

The basic principle behind the approach is very simple: the minimum weights are assigned to inputs and outputs of the DMU so that the ratio of the weighted output to weighted inputs of this unit (to be evaluated) is maximum, subject to the constraint that similar ratios of all the other DMUs of the group be equal to or less than 1. The decision variables are the weights (or multipliers) of the outputs (one for each output) and inputs (one for each input).

The mathematical format of envelopment analysis is stated as follows:

Objective function: Maximise: $$\frac{\sum_{j=1}^{M} \lambda_{Oj} y_{nj}}{\sum_{i=1}^{K} \theta_{O1} x_{ni}}$$

Subject to: (*a*) Output constraint: $$\frac{\sum_{j=1}^{M} \lambda_{Oj} y_{nj}}{\sum_{i=1}^{K} \theta_{O1} x_{ni}} \leq 1; \ (n = 1, 2, \ldots N)$$

(*b*) Non-negativity constraint: $\lambda_{oj}, \theta_{oi} \geq 0$ ($i = 1, 2, \ldots K$); ($j = 1, 2, \ldots M$)

In this model, N is the number of DMUs that convert K number of inputs to M number of outputs. It is assumed that (x_{ni}) are the inputs used by basic DMU (the one which is being evaluated) to produce (y_{nj}) outputs. λ and θ are weights to be determined by the model.

The model described in this format is called an output maximising multiplier version in DEA literature. It should also be noted that the model has been stated in non-linear fractional mathematical form. Its operational form is described in the following through Model 1.

The optimal value of the objective function is the efficiency score obtained by the basic DMU. In case the value of score is 1 (that is, 100 per cent efficient unit in relation to other units), the basic DMU is said to be located on the optimum scale of the production frontier. On the other hand, when score is less than 1, the basic DMU is located farther away from optimum scale frontier (detailed interpretation follows). Since the prices of inputs and outputs do not enter into this format of the model, allocative efficiency is not considered in this methodology.

The operational format of DEA model is in the duality form as described in Model 1:

Objective function: Minimise: $\theta_0 \lambda_0$ (Model 1)

Subject to constraints: (*a*) Output constraint: $\sum_{n=1}^{N} y_{nj}\lambda_n \ \geq y_{oj}\ (j = 1, 2,...M)$

(*b*) Input constraint: $\theta_0 x_{Oj} \geq \sum_{n=1}^{N} x_{ni}\lambda_n\ (i = 1, 2, ...N)$

(*c*) VRTS constraint: $\sum_{n=1}^{N} \lambda_n \leq 1$

(*d*) Non-negativity constraint: $\lambda_n \geq 0 \quad (n = 1, 2, \ldots N)$

The model states that for what minimum values of input-output weights (λ and θ):

1. Output of the reference set must be at least at the same level as the output of the basic unit;
2. Efficiency corrected input usage[1] of the basic unit must be greater than or the same as the input use of the reference set;[2] and
3. In case of VRTS constraint efficiency scores are computed under two assumptions: CRTS and VRTS; when constraint (*c*) is dropped, the frontier assumption changes from VRTS to CRTS.

In case the efficiency score of the DMU calculates to one with assumption (*c*), the basic DMU is said to be located on the VRTS frontier, otherwise it is located farther away from it (detailed interpretation follows).

The summation format of the operational model (described under Model 1) can also be transformed into matrix notations.

Assume that it is required to investigate the performance of nine hospitals, which use four identical inputs (number of beds, doctors, equipment and level of infrastructure) to produce three identical products (number of out-patients, in-patients and laboratory tests).

Therefore, y_{53} would describe outputs of the fifth hospital using x_{54} inputs; y_{83} would describe output of the eighth hospital using x_{83} inputs.

Supposing we wish to obtain the efficiency score of the ninth hospital in relation to remaining eight hospitals, the various equations of the model can then be written as:[3]

$$\begin{bmatrix} \lambda_1 y_{11} + \lambda_2 y_{21} + \ldots + \lambda_8 y_{81} \\ \lambda_1 y_{12} + \lambda_2 y_{22} + \ldots + \lambda_8 y_{82} \\ \lambda_1 y_{13} + \lambda_2 y_{23} + \ldots + \lambda_8 y_{83} \end{bmatrix} \geq \begin{bmatrix} y_{91} \\ y_{92} \\ y_{93} \end{bmatrix}_{3\times 1}$$

(*a*) Output constraint:

$$\begin{bmatrix} y_{11} & y_{21} & \cdots\cdots & y_{81} \\ y_{12} & y_{22} & \cdots\cdots & y_{82} \\ y_{13} & y_{23} & \cdots\cdots & y_{83} \end{bmatrix}_{3\times 8} \begin{bmatrix} \lambda_1 \\ \lambda_2 \\ \vdots \\ \lambda_8 \end{bmatrix}_{8\times 1} \geq \begin{bmatrix} y_{91} \\ y_{92} \\ y_{93} \end{bmatrix}_{3\times 1}$$

$$\mathbf{Yj} \qquad \boldsymbol{\lambda}'_{\mathbf{oj}} \geq \mathbf{y}_{\mathbf{oj}}$$

(*b*) Input constraint:

$$\begin{bmatrix} \theta x_{91} \\ \theta x_{91} \\ \theta x_{93} \\ \theta x_{94} \end{bmatrix}_{4\times 1} \geq \begin{bmatrix} \lambda_1 x_{11} + \lambda_2 x_{21} + \ldots + \lambda_8 x_{81} \\ \lambda_1 x_{12} + \lambda_2 x_{22} + \ldots + \lambda_8 x_{82} \\ \lambda_1 x_{13} + \lambda_2 x_{23} + \ldots + \lambda_8 x_{83} \\ \lambda_1 x_{14} + \lambda_2 x_{24} + \ldots + \lambda_8 x_{84} \end{bmatrix}_{4\times 1}$$

$$\boldsymbol{\theta}_0\, x_{0\dot{c}} \geq x_{\dot{c}}\ \boldsymbol{\lambda}'_{oj}$$

(*c*) Non-negativity constraint: $\lambda_{oj} \geq 0$

Objective function: $\theta_0 \lambda_{oj} = \boldsymbol{\theta}_0 \boldsymbol{\lambda}_{oj}\, \mathbf{I}'$

Note that since in objective function a single number of efficiency score is obtained, θ_0 is a scalar, λ_0 is a row vector (1×8) and *I*–is the identity column vector of (8×1).

Model I described in summation form can now be stated in the matrix notations in the following procedure:

Objective function:		Minimise:	$\boldsymbol{\theta}_0\, \boldsymbol{\lambda}_{0j}\, \mathbf{I}'$
Under the constraints:	(*a*)	Output constraint	$\mathbf{Y}_j\, \boldsymbol{\lambda}'_{0j} \geq \mathbf{Y}_{0j}$
	(*b*)	Input constraint	$\boldsymbol{\theta}_0\, \mathbf{x}_{0i} \geq \mathbf{Xi}\ \boldsymbol{\lambda}'_{0j}$
	(*c*)	VRTS constraint	$\boldsymbol{\lambda}_{0j}\, \mathbf{I}' \leq \mathbf{1}$
	(*d*)	Non-negativity constraint	$\boldsymbol{\lambda}_{0j} \geq \mathbf{0}$

Interpretation of Computed Efficiency Scores

As discussed earlier, relative efficiency scores are computed under two assumptions: CRTS and VRTS. The interpretation of the scores, therefore, needs to be understood separately.

The following two situations may arise when efficiency score is obtained under the CRTS assumption:

Interpretation of Efficiency Score When it is Equal to 1 DMU is located on point *E* in Figure 3(a), and is on CRTS frontier *OS*. It is said to be operating on optimum scale of production. It is technically and scale-wise an efficient DMU.

Interpretation of Efficiency Score When it is Less Than 1 DMU is not on the CRTS frontier, hence not operating on optimum scale of production. Any of the two situations may arise: (*a*) the DMU is at point *M* of the VRTS frontier, or (*b*) it is at point *N* as shown in Figure 3(b).

In case of (*a*), DMU is said to be inefficient scale-wise and the calculated score measures its scale inefficiency only. It stays technically efficient with reference to the VRTS frontier.

In case of situation (*b*), the DMU is said to be technically inefficient. The computed score measures its overall technical inefficiency depicted by

Figure 3
(a) Efficiency Score Equal to 1

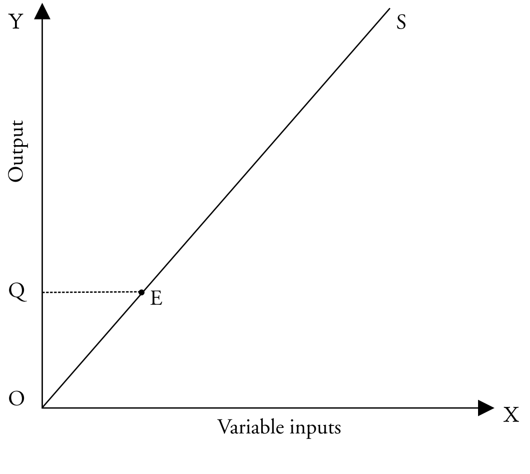
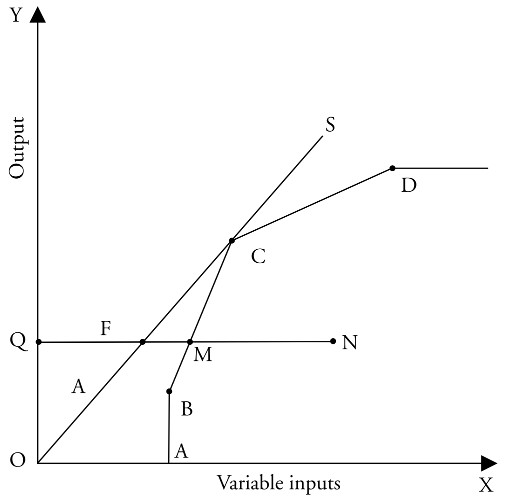


Figure 3
(b) Efficiency Score Less Than 1



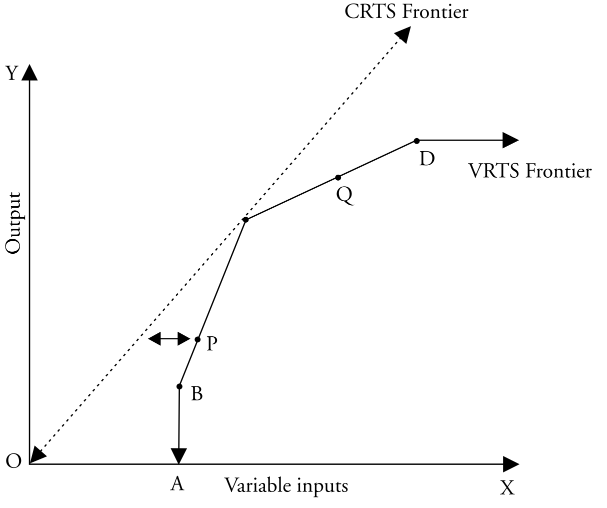
distance *EN* in Figure 3(b). The overall technical inefficiency is composite of scale inefficiency (= *EM*) and pure technical inefficiency (= *MN*).[4]

The following two situations may arise when efficiency score is computed under the VRTS assumption:

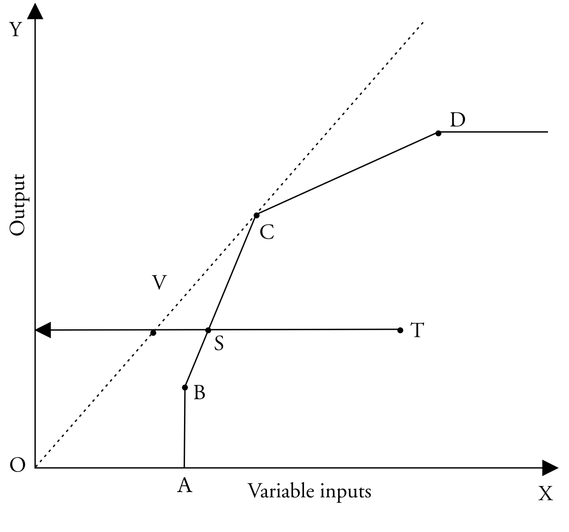
Interpretation of Efficiency Score When it is Equal to 1 The DMU may then be on either point *P* or *Q* of the VRTS frontier as shown in Figure 4(a), and experiencing variable returns to scale on *P* (*IRS*) and on *Q* (*DRS*). The DMU is then said to be only scale-wise inefficient; nevertheless, it is efficient technically. Hence, the computed efficiency score (= 1) reflects only technical soundness of the DMU with scale inefficiency equal to its distance from CRTS (marked by arrow).

Interpretation of Efficiency Score When it is Less than 1 The DMU is neither on the CRTS frontier nor on the VRTS frontier; it is farther away from both frontiers at point *T*, as shown in Figure 4(b). As such, the DMU is both technically and scale-wise inefficient. The computed efficiency score (less than 1) describes its pure technical inefficiency, which is equal to distance of *T* from the VRTS frontier (= *ST*). The scale inefficiency is obtained by distance from the CRTS frontier (=*VS*).

Figure 4
(a) Efficiency Score Equal to 1



(b) Efficiency Score Less Than 1



DEA: A Holistic Efficiency Measure for Health Services

Health-oriented services are multidimensional in nature. A single ratio or indicator (doctors per 100 beds; hi-tech equipment per 100 beds; etc.) cannot provide enough information of various dimensions of any health service. One is likely to derive erroneous conclusions from such single indicators and make normative benchmark comparisons. Again, ratio analysis evaluates output of health service in relation to input(s) consumed, thereby suppressing completely the implicit interdependency amongst the different services and resources. In fact, since providers of health care services operate in similar markets, resources employed may be considered immaterial (or homogenous) for all service providers, after adjusting for the scale of operation. Against this, DEA produces a single multidimensional score of performance by considering explicitly the mix of all the services provided through all the resources used.

Single efficiency score has other major advantages, which makes the envelope analysis more purposeful and useful. Single efficiency scores allow us to rank all the service providers as per their performance score, less efficient providers can then locate their position vis-à-vis their 'peers' who are able to use resources more efficiently. Second, less efficient providers can also identify the areas of improvement in the operation through this analysis.

Above all, in DEA, data on inputs and outputs needs no standardisation. Researchers can use data in the original units in which they are described or collected. They can also consider any number or kind of inputs and outputs of their own interest of the service providers under consideration. Finally this analysis works fine with a small sample size.

Through this article we have tried to uncover various principles and concepts underlying the non-parametric approach followed in DEA. The methodology explained here is static in nature; in other words, it measures the performance of service providers only at a point of time. A dynamic technique called Window Analysis suggested by Charnes et al. (1985) investigates variations of performance over the period of time. In the next issue of this journal we propose to provide application of static and dynamic approaches to evaluate the performance of public and private hospitals of a state.

Notes

1. Since the input correction factor is same for all types of inputs, the reduction in the observed inputs is proportional.
2. Reference set is set of all the remaining DMUs with which the basic unit is being evaluated.
3. A complete DEA model, therefore, involves solving nine such programmes each for a basic DMU ($n = 1$ to N) to get the efficiency scores of all DMUs. In each programme objective function and the input-output constraints will change accordingly.
4. Scale efficiency score is given by the ratio of efficiency score under CRTS to efficiency score under VRTS.

References

Banker, R., A. Charnes and **W. Cooper** (1984). Some models for estimating technical and scale inefficiencies in data envelopment analysis. *Management Science*, 30 (9), 1078–92.

Charnes, A., C.T. Clark, W.W. Cooper and **B. Golany** (1985). A development study of data envelopment analysis in measuring the efficiency of maintenance units in the US Air Forces. *Annals of Operation Research*, 2(1), 95–112.

Charnes, A., W.W. Cooper and **E. Rhodes** (1978). Measuring of efficiency of decision making units. *European Journal of Operational Research*, 2(6), 429–24.

Farrell, M.J. (1957). The measurement of productive efficiency. *Journal of Royal Statistical Society*, Series A, 120(3), 253–90.

P.R. Sodani is Associate Professor, Indian Institute of Health Management Research, 1 Prabhu Dayal Marg, Sanganer Airport, Jaipur 302011. E-mail: prsodani@yahoo.com

G.M.K. Madnani is Professor Emeritus of Management, Pacific Institute of Management, Post Box No.12, Pratap Nagar Extension, Airport Road, Udaipur, Rajasthan.