University of
Connecticut

Department of Economics Working Paper Series

Data Envelopment Analysis: An Overview

Subhash Ray
University of Connecticut

Working Paper 2014-33

November 2014

365 Fairfield Way, Unit 1063
Storrs, CT 06269-1063
Phone: (860) 486-3022
Fax: (860) 486-4463
http://www.econ.uconn.edu/

Data Envelopment Analysis: An Overview

Subhash C Ray
University of Connecticut
Storrs CT 06269-1063 USA
subhash.ray@uconn.edu

Abstract:

Over the past decades Data Envelopment Analysis (DEA) has emerged as an important nonparametric method of evaluating performance of decision making units through benchmarking. Although developed primarily for measuring technical efficiency, DEA is now applied extensively for measuring scale efficiency, cost efficiency, and profit efficiency as well. This paper integrates the different DEA models commonly applied in empirical research with their underlying theoretical foundations in neoclassical production economics.

JEL Classification Codes: C6, D2

Keywords: Linear Programming; Technical Efficiency; Returns to Scale; Distance Functions

November 2014

Data Envelopment Analysis: An Overview

Subhash C Ray

University of Connecticut

Storrs CT 06269-1063 USA

subhash.ray@uconn.edu

October 29, 2014

1. Introduction

Data Envelopment Analysis (DEA) is a nonparametric method of evaluating efficiency in resource utilization. In this approach the efficiency of a decision making unit (DMU) is evaluated by comparing the outcome of the actual decision with what is deemed to be best achievable outcome. It is obvious that what constitutes the 'best' outcome depends on the objective of the decision maker on one hand, and on the set of alternatives from which a particular decision is selected, on the other. The nature of decision making determines what are the choice variables and what are the parameters in a given context. For example, the manager of a primary healthcare center in a rural area might be assigned a given amount of resources (in the form of physicians and supplies) and the objective would be to immunize as many infants as possible. In this case, efficiency would be measured by the ratio of the actual number of infants immunized and the maximum number possible given the resources. This is output-oriented technical efficiency. A different example would be one where a landscaper has to irrigate a lawn of a given size and efficiency lies in getting the task completed using the minimum necessary amount of water. This relates to input-oriented technical efficiency. In the output-oriented case, realizing the full output potential is of primary importance. In the input-oriented case, conserving inputs has priority over expanding the output. In many other cases, there are market prices of inputs reflecting the relative worth of individual inputs. In such cases, the objective may be to produce the target output at the minimum cost. This may actually involve increasing the quantity used of some input so long as the resulting increase in cost is more than offset by the saving resulting from economizing on the use of a more valuable input. It may be noted that cost minimization is a valid objective even for public sector and other non-profit agencies (like schools and hospitals) because any cost saving ultimately releases valuable resources for the production of other outputs. Finally, from the perspective of a business enterprise engaged in producing outputs for profit, the bottom line is the amount of profit earned. In this case, all outputs and all inputs are choice variables and the only constraint on the producer's behavior is that the input-output bundle selected must be such that it must be technologically possible to produce the planned output from the input bundle selected. There are economic theories of producer's behavior corresponding to the alternative objectives. Correspondingly, there are appropriate DEA optimization problems that yield the relevant benchmarks for comparison with the actual outcome for evaluating efficiency.

The paper is organized as follows. Section 2 introduces the concept of the production technology and defines measures of output and input oriented technical efficiency. The corresponding DEA LP models are also formulated. Section 3 considers in details returns to scale properties of the technology and how to measure scale efficiency. The most productive scale size (MPSS) for a given input-output bundle is define and alternative ways to identify the nature of returns to scale at a particular point on the efficient frontier of the technology set are described. Section 4 covers non-radial measures of technical efficiency. Section 5 explains the concepts of the Distance Function, the Directional Distance Function, and the Geometric Distance Function. Section 6 deals with the question of invariance of different efficiency measures in light of data transformation. Section 7 provides measures of technical efficiency in the presence of bad or undesirable outputs. Section 8 deals with measurement of cost and profit efficiency. Section 9 is the conclusion.

2. The Production Technology

At the core of productivity and efficiency analysis is the concept of the production technology described by the production possibility set. Production is the process of converting inputs into outputs. A bundle of inputs $(x \in R_n^+)$ is acquired by the producer from outside. It then goes through various parallel or sequential processes of transformation and ultimately exits the jurisdiction of the firm as a finished product in the form of an output bundle $(y \in R_m^+)$. A pair of input-output bundles (*x*, *y*) is a feasible production plan if the output bundle *y* can be produced from the input bundle *x*. The production possible set (*T*) includes all feasible production plans. Thus,

$$T = \{(x, y) : x \text{ can produce } y\}. \qquad (1)$$

Often the production possibility set is defined by means of a *production correspondence* or a *transformation function*

$$F(x, y) = \alpha \qquad (2)$$

mapping from R^{+}_{m+n} to the (0,1) interval on the real line. The production possibility set then can be expressed as:

$$T = \{(x, y): F(x, y) \le 1\} \qquad (3)$$

It is assumed that the production correspondence is non-increasing in inputs and non-decreasing in outputs. Thus, $F_i = \frac{\partial F}{\partial x_i} \le 0$ for each input *i* and $F_j = \frac{\partial F}{\partial y_j} \ge 0$ for each output *j*. These are also known as free disposability assumptions. They imply that if any input-output bundle is feasible, increase in any input not accompanied by a decrease in another input or increase in any output will not render the new input-output bundle infeasible. This rules out negative marginal productivity. Similarly, a decrease in any output quantity will not affect feasibility. Moreover if $F(x^0, y^0) = 1,\ (x^0, y^0)$ is a technically efficient input-output bundle

In the single output case, one uses the *production function*

$$y^* = f(x) \qquad (4)$$

where y^* is the maximum quantity of the scalar output that can be produced from the input bundle, *x*. In this case,

$$T = \{(x, y) : y \le f(x)\}. \qquad (5)$$

In parametric analysis, one specifies an explicit form of the production function and uses statistical estimation techniques like the maximum likelihood procedure to calibrate the parameters of the specified function using sample data of inputs and outputs[1]. In Data Envelopment Analysis one avoids any kind of functional specification and instead makes a number of quite general assumptions about the nature of the underlying production technology to construct the production possibility set from sample data.

[1] The stochastic production function was introduced by Aigner, Lovell, and Schmidt (1977) and Meeusen, W. and J. van den Broeck (1977). For an excellent and comprehensive exposition of Stochastic Frontier Analysis (SFA) see Kumbhakar and Lovell (2000)

Let the data set $D = \left\{(x^j, y^j);\ j = 1, 2, ..., N\right\}$ be the set of observed input-output bundles of *N* firms from a particular industry. The following assumptions are made about the production technology.

(a) Each observed input-output bundle is feasible.

(b) The production possibility set is convex.

(c) Inputs are freely disposable. That is if $(x^0, y^0) \in T$ and $x^1 \geq x^0$, then $(x^1, y^0) \in T$.

(d) Outputs are freely disposable. That is if $(x^0, y^0) \in T$ and $y^1 \leq y^0$, then $(x^0, y^1) \in T$.

There would, of course, be infinitely many sets satisfying these assumptions. In DEA *T* is estimated by the set

$$S = \left\{(x, y) : x \geq \sum_{j=1}^{N} \lambda_j x^j;\ y \leq \sum_{j=1}^{N} \lambda_j y^j;\ \sum_{j=1}^{N} \lambda_j = 1; \lambda_j \geq 0; (j = 1, 2, ..., N)\right\}. \qquad (6)$$

It is the smallest convex set containing the observed data points and satisfying the free disposability assumption. It is also known as the free disposal convex hull of the set *D*.

Technical Efficiency

The *output-oriented* technical efficiency[2] of a firm producing the output bundle y^0 from the input bundle x^0 is measured as

$$\tau_y(x^0, y^0) = \frac{1}{\varphi^*} \qquad (7)$$

where

$$\begin{aligned} \varphi^* &= \max \varphi \\ s.t. &\sum_{j=1}^{N} \lambda_j y^j \geq \varphi y^0; \\ &\sum_{j=1}^{N} \lambda_j x^j \leq x^0; \qquad (8) \\ &\sum_{j=1}^{N} \lambda_j = 1; \\ &\lambda_j \geq 0; (j = 1, 2, ..., N);\ \varphi\, unrestricted. \end{aligned}$$

[2] This is also known as Farrell efficiency named after Farrell (1957) who extended the earlier work by Debreu (1951) and Shephard (1953). However, the multiple output linear programming formulation is due to Charnes, Cooper, and Rhodes (CCR) (1978). The model in (8) is a generalization of the original CCR model by Banker, Charnes, and Cooper (BCC) (1984) that allows variable returns to scale.

Note by convexity $\sum_{j=1}^{N}\lambda_j = 1$ and $\lambda_j \ge 0$ ensures that the input-output bundle $\left(\sum_{j=1}^{N}\lambda_j x^j, \sum_{j=1}^{N}\lambda_j x^j\right)$ is feasible. Due to free disposability of inputs $x^0 \ge \sum_{j=1}^{N}\lambda_j x^j \Rightarrow \left(x^0, \sum_{j=1}^{N}\lambda_j y^j\right)$ is also feasible. Finally, due to free disposability of outputs $\varphi y^0 \le \sum_{j=1}^{N}\lambda_j x^j \Rightarrow \left(x^0, \varphi \mathrm{y}^0\right)$ is feasible. The optimal value of the objective function in (8) shows the maximum rate by which *all outputs* of the firm can be expanded without any increase in any individual input. When different outputs can be expanded at different rates, φ^* is the lowest of these expansion factors. For example, in a 2-output case, if one output can be expanded by a factor of 1.5 and the other by a factor of 1.25, φ^* equals the lower of the two values. In this case, it is possible to expand the *output bundle* itself by at least 25% across the board one output can be expanded even beyond that. The output-oriented technical efficiency is 0.80 implying that it is realizing only 80% of the potential output producible from its current input bundle.

An alternative measure of technical efficiency of the firm is its *input-oriented* technical efficiency

$$\begin{aligned}
&\tau_x(x^0, y^0) = \min \theta \\
&s.t. \sum_{j=1}^{N}\lambda_j y^j \ge y^0; \\
&\quad \sum_{j=1}^{N}\lambda_j x^j \le \theta x^0; \qquad\qquad (9) \\
&\quad \sum_{j=1}^{N}\lambda_j = 1; \\
&\lambda_j \ge 0; (j = 1, 2, ..., N);\ \theta\, unrestricted.
\end{aligned}$$

The input-oriented technical efficiency of the firm shows the factor by which the entire input bundle can be scaled down without reducing any output. In the multiple input case it may be possible to reduce individual inputs even further. In general, the input- and output-oriented technical efficiency measures of a firm will be different.

When the production possibility set is defined by a transformation function or production correspondence as in (3), the *graph* of the technology is

$$G = \{(x, y): F(x, y) = 1\} \qquad (10)$$

The nonparametric version of the graph would be

$$G = \left\{(x, y) : (x, y) \in T; \beta < 1 \Rightarrow (\beta x, y) \notin T; \alpha > 1 \Rightarrow (x, \alpha y) \notin T\right\}. \qquad (11)$$

It is apparent that every $(x, y) \in G$ is technically efficient in both the input and the output orientation. The graph of the technology constitutes the frontier of the production possibility set.

Constant Returns to Scale

The technology exhibits constant returns to scale (CRS) globally if

$$(x, y) \in T \Rightarrow (kx, ky) \in T \ \forall \ k \geq 0.$$

An implication of CRS is that any non-negative radial expansion or contraction of a feasible input-output bundle is also a feasible input-output bundle. Under the CRS assumption, an empirical estimate of the production possibility set is

$$S^C = \left\{(x, y) : x \geq \sum_{j=1}^{N} \lambda_j x^j ; y \leq \sum_{j=1}^{N} \lambda_j y^j ; \lambda_j \geq 0; (j = 1, 2, ..., N)\right\}. \qquad (12)$$

Note the absence of the restriction that the λs add up to unity. When only convexity is assumed all weighted averages of observed input output bundles are also feasible. It was necessary that the weights add up to 1. Thus so long as $\sum_{j=1}^{N} \lambda_j = 1$ and each λ_j is non-negative, $(\sum_{j=1}^{N} \lambda_j x^j, \sum_{j=1}^{N} \lambda_j y^j)$ is feasible. But with the added assumption of CRS, $(k\sum_{j=1}^{N} \lambda_j x^j, k\sum_{j=1}^{N} \lambda_j y^j)$ is also feasible for any $k \geq 0$. Now consider the weights $\mu_j = k\lambda_j, k \geq 0.$ CRS implies that $(\sum_{j=1}^{N} \mu_j x^j, \sum_{j=1}^{N} \mu_j y^j)$ is feasible. But $\sum_{j=1}^{N} \mu_j = k$ which can be any non-negative number. This explains why the weights need not add up to 1 in (12).

The set S^C is sometimes described as the free disposal conical hull of *D*.

The output-oriented CRS technical efficiency[3] is

$$\tau_y^C(x^0, y^0) = \frac{1}{\varphi_C^*} \qquad (13)$$

where

$$\begin{aligned} &\varphi_C^* = \max \varphi \\ &s.t. \sum_{j=1}^{N} \lambda_j y^j \geq \varphi y^0; \\ &\quad \sum_{j=1}^{N} \lambda_j x^j \leq x^0; \\ &\quad \lambda_j \geq 0; (j = 1, 2, ..., N);\ \varphi\, unrestricted. \end{aligned} \qquad (14)$$

Similarly, the input-oriented CRS technical efficiency is

$$\begin{aligned} &\tau_x^C(x^0, y^0) = \min \theta \\ &s.t. \sum_{j=1}^{N} \lambda_j y^j \geq y^0; \\ &\quad \sum_{j=1}^{N} \lambda_j x^j \leq \theta x^0; \\ &\quad \lambda_j \geq 0; (j = 1, 2, ..., N);\ \theta\, unrestricted. \end{aligned} \qquad (15)$$

It is easy to verify that under the CRS assumption input- and output-oriented measures of technical efficiency are identical.

Figures 1(a-b) explain graphically the concepts of output and input-oriented technical efficiency for variable and constant returns to scale in the 1-input 1-output case. In Figure 1a the curve y^*= *f(x)* represents the production function. Point *A* represents in the input-output combination (x_A, y_A). The maximum output producible from input x_A is $y_A^* = f(x_A)$. The efficient input-output bundle (x_A, y_A^*) is shown by the point A^* on the frontier. The output-oriented technical efficiency of the firm *A* is measured as $\tau_y^A = \frac{y_A}{y_A^*} = \frac{Ax_A}{A^* x_A}$. Similarly, for firm *B*, the actual output produced from input x_B is y_B and the maximum producible is y_B^*. The output-oriented technical efficiency of

[3] This is the CCR model.

firm *B* is $\tau_y^B = \frac{y_B}{y_B^*} = \frac{Bx_B}{B^* x_B}$. Further, the minimum quantity of the input needed for producing output y_A is $x_A^* = f^{-1}(y_A)$. Hence, the input-oriented technical efficiency of *A* is $\tau_x^A = \frac{x_A^*}{x_A} = \frac{Cy_A}{Ay_A}$. The corresponding input-oriented technical efficiency of *B* is $\tau_x^B = \frac{x_B^*}{x_B} = \frac{Dy_B}{By_B}$. It is clear that, in general, the input- and output-oriented technical efficiencies of the same firm will differ.

Figure 1b illustrates the case of CRS. Here the production function is *f(x)*= *kx*. It is clear from the properties of similar triangles that $\frac{Ox^*}{Ox_0} = \frac{OC}{OA^*} = \frac{Ax_0}{A^* x_0}$. Hence, $\tau_x = \tau_y$ for every firm.

The dual of the maximization problem in (14) is[4]

$$\begin{aligned} &\min u^{0}{}'x^{0} \\ &s.t. u^{0}{}'x^{j} - v^{0}{}'y^{j} \geq 0; (j = 1, 2, ..., N); \\ &v^{0}{}'y^{0} = 1; \\ &u^{0}, v^{0} \geq 0. \end{aligned} \qquad (16)$$

Several points need special attention in this problem. First, x^j is the observed input vector and y^j the corresponding output vector of firm *j*. These are all parameters. Also, (x^0, y^0) is one of the (x^j, y^j) bundles. The vectors u^0 and v^0 are the choice variables in this problem. These can be interpreted as the *shadow prices* of the inputs and outputs. Second, the optimal values of these dual variables depend upon the values of (x^0, y^0). That is the reason why they are superscripted. In (16) these shadow prices of inputs and outputs are chosen in such a way that evaluated at these prices

(a) the shadow value of the output bundle y^0 equals unity and

(b) the shadow value of any observed output bundle cannot exceed the shadow cost of the corresponding input bundle. This is true for the (x^0, y^0) bundle as well.

The problem in (16) can easily be recast as the original linear fractional functional programming problem:

[4] This is the *multiplier form* of the model.

$$\begin{aligned}&\min \tfrac{u^{0}{}'x^{0}}{v^{0}{}'y^{0}}\\&s.t. \tfrac{u^{0}{}'x^{j}}{v^{0}{}'y^{j}} \geq 1; (j=1,2,...,N); \qquad (17)\\&u^{0}, v^{0} \geq 0.\end{aligned}$$

This can also be expressed as

$$\begin{aligned}&\max \tfrac{v^{0}{}'y^{0}}{u^{0}{}'x^{0}}\\&s.t. \tfrac{v^{0}{}'y^{j}}{u^{0}{}'x^{j}} \leq 1; (j=1,2,...,N); \qquad (18)\\&u^{0}, v^{0} \geq 0.\end{aligned}$$

This is the so called *ratio form* of the DEA problem introduced by Charnes, Cooper, and Rhodes (CCR) in their pioneering 1978 paper[5]. The LP problems in (14) and (15) above are generally called the CCR output- and input-oriented- DEA models. By contract the previous models in (8) or (9) are corresponding BCC models named after Banker, Charnes, and Cooper (1984). The CCR models are known as the CRS problems. Because no specific assumption is made about returns to scale in the BCC models, they are described as variable returns to scale (or VRS) problems.

The Multipliers as Shadow Prices

It may be noted that the inverse of the output-oriented proportional expansion factor in (14) is clearly a technical efficiency measure. By contrast, the ratio measure in (18) is a total productivity measure. It is true that by standard duality results they can be shown to be mathematically equivalent. But as noted by Førsund (2013) they are conceptually quite different. The Farrell efficiency measure relates directly to the frontier of the production possibility set. But there is no such obvious link in the case of the ratio measure. This is one reason why there is so much confusion about the interpretation of the aggregation weights in the CCR model in its multiplier form in the OR literature and people have tried to impose arbitrary weight restrictions to avoid zero weights. However, as shown below that the weights come from the gradient of a

[5] They use a normalization that was fist considered in Charnes and Cooper (1968).

supporting hyperplane at the efficient projection of an observed input-output bundle and, contrary to the popular belief, are far from arbitrary. As shown by Førsund (2013), the *ratios* of the multipliers associated with two inputs show the marginal rate of substitution between those inputs. Similarly, for a pair of outputs the ratio of the multipliers show the marginal rate of transformation between them.

Slacks

An inherent problem with inequality-constrained optimization is that more often than some of the constraints prove to be non-binding at the optimal solution. This results in the presence of slacks in the input and/or the output constraints. As such, presence of positive slacks at the optimal solution of an LP problem poses no particular problem. After all, when constraints are in the form of weak inequalities, it is only common to end up with slacks. In DEA, however, presence of slacks in the output or input constraints has been a matter of concern right from the start. This is because of what they imply about of the measured technical efficiency of the firm. Consider, for example, the following 2-input 1-output BCC input-oriented problem:

$$\begin{aligned}
&\min \theta \\
&s.t. \sum_{j=1}^{N} \lambda_j x_{1j} \le \theta x_{10}; \\
&\sum_{j=1}^{N} \lambda_j x_{2j} \le \theta x_{20}; \\
&\sum_{j=1}^{N} \lambda_j y_{2j} \ge y_0; \\
&\sum_{j=1}^{N} \lambda_j = 1; \lambda_j \ge 0; (j = 1, 2, ..., N).
\end{aligned} \qquad (19)$$

Now suppose that at the optimal solution, $\sum_{j=1}^{N} \lambda_j^* x_{1j} = x_{10}$ but $\sum_{j=1}^{N} \lambda_j^* x_{2j} = 0.5 x_{20}$. Obviously, in this case θ^* equals unity and the firm under evaluation is considered to be operating at 100% efficiency. This, of course, is quite difficult to accept given the fact that it can cut down its use of input 2 by half without reducing output or increasing input 1. In order to correct this anomaly CCR (1979) modified their original formulation of the problem so that a firm could be considered efficient only when θ^* (or φ^*, when it is an output-oriented model) was 1 *as well as all input and output slacks were 0.* In order to enforce this added requirement, the included a very small penalty (ε) for the presence of any output or input slack in the objective function. A revised version of (19) incorporating the slacks explicitly would be:

$$\begin{aligned}
&\min \theta - \varepsilon[s_1^- + s_2^- + s_1^+] \\
&s.t. \sum_{j=1}^{N} \lambda_j x_{1j} + s_1^- = \theta x_{10}; \\
&\quad \sum_{j=1}^{N} \lambda_j x_{2j} + s_2^- = \theta x_{20}; \\
&\quad \sum_{j=1}^{N} \lambda_j y_{2j} - s_1^+ = y_0; \qquad\qquad (20) \\
&\quad \sum_{j=1}^{N} \lambda_j = 1; \\
&\quad s_1^-,\, s_2^-,\, s_1^+ \geq 0; \\
&\quad \lambda_j \geq 0; (j = 1, 2, ..., N).
\end{aligned}$$

Here (s_1^-, s_2^-) are the input slacks and s_1^+ is the output slack. CCR stipulated that ε should be a *non-Archemdian* (or *infinitesimal)* positive number in order to ensure that decreasing θ gets a preemptive priority over increasing input or output slacks in the solution algorithm of the problem. But a practical question is: what numerical value should one use to actually solve the problem in (20)? No matter how small a value one can choose, it is always possible to pick one that is smaller. Thus, the minimization problem in (20) cannot be actually solved. It can be seen, however, that irrespective of any numerical value of ε, the objective first is to minimize θ and then to maximize the sum of the slacks. In practice the problem is solved in two steps. In step 1, one solves the problem in (19) without any concern about slacks. If there are multiple optimal solutions, there will be different vectors λ^* going with the same minimum value θ^*. The objective in the second stage is to select the optimal solution that maximizes the sum of the slacks through the following model:

$$
\begin{aligned}
&\max s_1^{-1} + s_2^{-} + s_1^{+} \\
&s.t. \sum_{j=1}^{N} \lambda_j x_{1j} + s_1^{-} = \theta^{*} x_{10}; \\
&\sum_{j=1}^{N} \lambda_j x_{2j} + s_2^{-} = \theta^{*} x_{20}; \\
&\sum_{j=1}^{N} \lambda_j y_{2j} - s_1^{+} = y_0; \qquad\qquad (21) \\
&\sum_{j=1}^{N} \lambda_j = 1; \\
&s_1^{-},\, s_2^{-},\, s_1^{+} \geq 0; \\
&\lambda_j \geq 0; (j = 1, 2, ..., N).
\end{aligned}
$$

It is important to note that in this second step problem θ^* is a parameter rather than a choice variable. The optimal value of θ obtained from (19) along with those of the slacks obtained from s(21) constitute the optimal solution for (20) irrespective of an actual numerical value of ε. The practical usefulness of this 2-step procedure is not very clear. It certainly flags the presence of slacks at an optimal solution of (19) even when the optimal value of θ is 1. But it does not provide a more comprehensive measure of efficiency that incorporates penalties for the presence of slacks. The optimal values of the λs in (19) define the technically efficient projection $(x_0^* = \sum_{j=1}^{N} \lambda_j^* x^j, y_0^* = \sum_{j=1}^{N} \lambda_j^* y^j)$ of the input-output bundle (x^0, y^0). When multiple optimal solutions exist for (19), there are many such projections. The problem in step 2 helps to select among them. The CCR method provides no justification for the criterion of choice behind the problem in (21).

An alternative to (20) would be the so called *additive model* that simply maximizes the sum of the input and output slacks and is formulated as follows:

$$
\begin{aligned}
&\max \sum_{r=1}^{m} s_r^{+} + \sum_{i=1}^{n} s_i^{-} \\
&s.t. \quad \sum_{j=1}^{N} \lambda_j y_{rj} - s_r^{+} = y_{r0}; (r = 1,2,...,m); \\
&\qquad \sum_{j=1}^{N} \lambda_j x_{ij} + s_i^{-} = x_{i0}; (i = 1,2,...,n); \\
&\qquad \sum_{j=1}^{N} \lambda_j = 1; \qquad\qquad\qquad\qquad (22) \\
&\quad s_i^{-} \geq 0, (i = 1,2,...,n); \\
&\quad s_r^{+} \geq 0, (r = 1,2,...,m); \\
&\quad \lambda_j \geq 0; (j = 1,2,...,N).
\end{aligned}
$$

Although sometimes used in the literature, this additive model is useless as a measure of efficiency. The objective function is the sum of input and output slacks that are measured in heterogeneous units and has no meaning. Its only usefulness lies in the fact that this will be 0 only when the θ^* in (20) (or φ^* in an output-oriented model) equals unity while all input and output slacks are 0.

3. Scale Efficiency

While full technical efficiency requires a firm to produce the maximum output(s) from its observed input bundle, in order to be considered scale efficient the firm needs to operate at the scale where average productivity reaches a maximum. Because average productivity is a meaningful concept only when a single output is produced from a single input, the concept of scale efficiency is best described in the context of a 1-input 1-output technology. Consider a firm with input-output (x_0, y_0). Its average productivity is $\frac{y_0}{x_0}$. Clearly, if it is not technically efficient, it is possibly to increase the output without changing the input or to lower the input without reducing the output. In either case, its productivity would increase. Now suppose that the production function is $y^* = f(x)$ and the corresponding graph of the technology is

$$G = \{(x, y) : y = f(x); x \geq 0; y \geq 0\}. \qquad (23)$$

For any $(x, y) \in G, AP(x) = \frac{y}{x} = \frac{f(x)}{x}$. Thus, if $(x_0, y_0) \in G, AP(x_0) = \frac{f(x_0)}{x_0}$. Because y_0 is the maximum output producible from input x_0 an increase in average productivity is not possible so long as the input level does not change. There may exist other input levels, however, where the average productivity is higher. Let x^*be the input level where average productivity attains a maximum. In that case, $\frac{dAP(x)}{dx}=\frac{xf'(x)-f(x)}{x^2}=0$ at the input level x^*. Frisch (1965) described the input level where average productivity is maximum as the *technical optimal production scale* (TOPS).

Several points are to be noted:

(a) At the technically optimal input level (x^*) locally constant returns to scale holds (because $\frac{dAP(x)}{dx}|_{x=x^*}=0.$

(b) $AP(x)=\frac{f(x)}{x}\leq AP(x^*)=\frac{f(x^*)}{x^*}$ for all input levels (x).

(c) At the input level (x^*) marginal productivity and average productivity are equal. Thus, $f'(x^*)=\frac{f(x^*)}{x^*}.$ This implies that $f(x^*)=f'(x^*).x^*.$

The scale efficiency of the firm operating at the input level *x0* is

$$SE(x_0)=\frac{AP(x_0)}{AP(x^*)}\leq 1. \qquad (24)$$

More specifically,

$$SE(x_0)=\frac{\frac{f(x_0)}{x_0}}{\frac{f(x^*)}{x^*}}=\frac{f(x_0)}{x_0 f'(x^*)}. \qquad (25)$$

Now define $\delta\equiv f'(x^*)$ and consider a *pseudo* production function

$$y^{**}=r(x)=\delta x. \qquad (26)$$

Then, the denominator in (25) becomes $\alpha x_0=r(x_0).$ Therefore, an alternative measure of scale efficiency is

$$SE(x_0)=\frac{f(x_0)}{r(x_0)}. \qquad (27)$$

Note that (b)and (c) above together imply

$$f(x)\leq f'(x^*)x=\delta x=r(x) \text{ and}$$

$$f(x^*)\leq f'(x^*)x^*=\delta x^*=r(x^*).$$

In other words, $f(x)\leq r(x)$ for all x and $f(x)=r(x)$ at $x = x^*$. In particular, $f(x_0)\leq r(x_0)$. This ensures that $SE(x_0)\leq 1$.

The *pseudo* production function is a tangent to the graph of the technology and *would have been* the production frontier if the technology did exhibit globally constant returns to scale. Note that under variable returns to scale, $y^* = f(x)$ is the true production function and the right measure of technical efficiency is $\tau_y = \frac{y_0}{y_0^*} = \frac{y_0}{f(x_0)}$. The other measure $\tau_y^C = \frac{y_0}{y_0^{**}} = \frac{y_0}{r(x_0)}$ is not the correct measure unless the technology exhibits CRS at all input levels. But this CRS efficiency measure, false as it is, does serve a useful purpose because the ratio of the CRS and VRS measures of technical efficiency is a valid measure of scale efficiency. This can be shown as follows:

$$SE\left(x_0\right) = \frac{f(x_0)}{r(x_0)}$$
$$= \frac{\frac{y_0}{r(x_0)}}{\frac{y_0}{f(x_0)}} = \frac{\tau_y^C(x_0,y_0)}{\tau_y(x_0,y_0)}. \qquad (28)$$

It should be noted that the expression in (28) measures the *output-oriented* scale efficiency of the input level *x₀*. In a perfectly analogous manner, one can take the output level *y₀* as given and measure the *input-oriented* scale efficiency

$$SE(y_0) = \frac{\tau_x^C(x_0,y_0)}{\tau_x(x_0,y_0)}. \qquad (29)$$

In Figure 2 the curve $y^* = f(x)$ is the true VRS production function. At the input level x_0 the firm produces output $y_0 < f(x_0)$. This input-output combination is shown by the point *A*. If technical inefficiency is eliminated, it could move to the point *B* on the production function. Here its average productivity would be $AP^*(x_0) = \frac{Bx_0}{Ox_0}$. But the maximum average productivity is along the production function is attained at the point *C* where the input level is x^* and this maximum average productivity is $AP^*(x^*) = \frac{Cx^*}{Ox^*}$. Scale efficiency at input level x_0 is

$SE(x_0) = \frac{AP^*(x_0)}{AP^*(x^*).}$ Now consider the tangent line $y^{**} = r(x)$. The point *D* shows the maximum output, y_0^{**} that would have been producible from input x_0 if CRS held. Now average productivity remains constant along the tangent line. Thus, comparing the average productivities at points *B* and *C* is equivalent to comparing productivities at points *B* and *D*. But average productivity at *D* would be $\frac{r(x_0)}{x_0}$ whereas at *B* it is $\frac{f(x_0)}{x_0}$. Thus, scale efficiency at input level x_0 is $\frac{f(x_0)}{r(x_0)} = \frac{Bx_0}{Dx_o}$.

Ray Average Productivity and Returns to Scale

The single input single output case was useful for illustrative purposes but is of little relevance in real life because seldom if ever any output is produced from one input alone. We now consider a multiple input single output technology. The production function now shows the maximum *scalar* output producible from a *vector* of inputs. Consider an input-output combination (x^0, y_0) that lies in the graph, *G*. That is,

$$y_0 = f(x^0);\ x^0 = (x_{10}, x_{20,}..., x_{n0}).$$

Now consider another bundle (x^1, y_1) also in the graph such that $x^1 = \beta x^0$.The two input bundles differ only in scale but not in input proportions. The vectors x^0 and x^1 lie on the same ray through the origin in the input space. If the bundle x^0 is considered to be 1 unit of a *composite input* then x^1 represents β units of the same input. If β>1,the bundle x^1 is a radial expansion of the x^0 bundle. Now suppose that $y_1 = \alpha y_{0.}$ The *ray average productivity* measured by *output per unit of the composite input* at (x^0, y_0) is y_0 and at (x^1, y_1) is $\frac{\alpha y_0}{\beta}$. If $\alpha > \beta > 1$, then *ray average productivity* is increasing ar (x^0, y_0) and we conclude that locally increasing returns to scale (IRS) holds at this point on the graph. On the other hand, $1 < \alpha < \beta$ signifies locally diminishing returns to scale (DRS). Finally, $\alpha = \beta$ implies constant returns to scale (CRS). Note that these are all local characteristics of the technology and are evaluated as $\beta \to 1$ from above. The technology may exhibit increasing, constant, or diminishing returns to scale at different points on the graph. This is why it is described as variable returns to scale (VRS).

Most Productive Scale Size

Banker (1984) generalized Frisch's concept of the technically optimal production scale to the multiple output multiple input case. A feasible input-output bundle (x^0, y^0) is a *most productive scale size* (MPSS) if for all nonnegative scalars (α, β) for which (βx^0, αy^0) is a feasible input-output combination, $\frac{\alpha}{\beta} \leq 1$. In other words, ($x^0$, y^0) is an MPSS only if there is no other feasible input-output bundle with the same mix of inputs and outputs but a higher *ray average productivity.* It is obvious that no feasible input-output bundle can be an MPSS unless it is in the graph. Recall that if $(x^0, y^0) \in T$ but $\notin G$, then there will exist either some $\beta < 1$ such that $(\beta x^0, y^0) \in T$ or some $\alpha > 1$ such that $(x^0, \alpha y^0) \in T$. In the former case, one gets $\frac{\alpha}{\beta} > 1$ for $\alpha = 1$. In the latter case, $\frac{\alpha}{\beta} > 1$ for $\beta = 1$.

The following lemma due to Ray (2009) shows that when the production possibility set is convex, IRS holds at all scales smaller than the smallest MPSS. Similarly, DRS holds at all scales larger than the largest MPSS.

Lemma: *For any convex productivity possibility set T, if there exist non-negative scalars α and β such that α >β >1, and both* $(\bar{x}, \bar{y})$ *and* $(\beta\bar{x},\ \alpha\bar{y}) \in G,$ *then* $\gamma > \delta$ *for every γ and δ such that 1<δ<β and* $(\delta\bar{x},\ \gamma\bar{y}) \in G.$

Proof: Because $(\bar{x}, \bar{y})$ and $(\beta\bar{x},\ \alpha\bar{y})$ are both feasible, by convexity of *T*, for every $\lambda \in (0,1), ((\lambda + (1-\lambda)\beta)\bar{x},\ (\lambda + (1-\lambda)\alpha)\bar{y})$ is also feasible. Now select λ such that $\lambda + (1-\lambda)\beta = \delta.$ Further, define $\mu = \lambda + (1-\lambda)\alpha.$ Using these notations, $(\delta\bar{x}, \mu\bar{y}) \in T.$ But, because $(\delta\bar{x},\ \gamma\bar{y}) \in G,\ \gamma \geq \mu.$ However, because $\alpha > \beta, \mu > \delta.$ Hence, $\gamma > \delta.$

An implication of this lemma is that, when the production possibility set is convex, if the technology exhibits locally diminishing returns to scale at smaller input scale, it cannot exhibit increasing returns at a bigger input scale. This is easily understood in the single-input single-output case. When both *x* and *y* are scalars, average productivity at $(\bar{x}, \bar{y})$ is $\frac{\bar{y}}{\bar{x}}$ and at $(\beta\bar{x}, \alpha\bar{y})$ it is $\frac{\alpha}{\beta}\frac{\bar{y}}{\bar{x}}.$ Thus, when $\alpha > \beta,$ average productivity has increased. The above lemma implies that for every input level *x* in between $\bar{x}$ and $\beta\bar{x}$, average productivity is greater than $\frac{\bar{y}}{\bar{x}}$. Thus, average productivity could not first decline and then increase as the input level increased from $\bar{x}$ to $\beta\bar{x}$.

Two results follow immediately. First, locally increasing returns to scale holds at every input-output bundle *(x, y)* $\in G$ that is smaller than the smallest MPSS. Second, locally diminishing returns to scale holds at every input-output bundle *(x, y)* $\in G$ that is greater than the largest MPSS. To see this, let $x = bx^*$ and $y = ay^*$, where (x^*, y^*) is the smallest MPSS for the given input and output mix. Because *(x, y)* is not an MPSS, $\frac{a}{b} < 1.$ Further, assume that *b* < 1. Define $\beta = \frac{1}{b}(>1)$ and $\alpha = \frac{1}{a}.$ Then $(x^*, y^*) = (\beta x, \alpha y)$ and $\frac{\alpha}{\beta} > 1.$ Because ray average productivity is higher at a larger input scale, by virtue of the lemma, locally increasing returns to scale holds at (*x*, y). Next assume that *b* > 1. Again, because *(x, y)* is not an MPSS, $\frac{a}{b}$ <1. That is ray average productivity has fallen as the input scale is increased from x^* to $x = bx^*$. Then, by virtue of the lemma, ray average product could not be any higher than $\frac{a}{b}$ at a slightly greater input scale, $\bar{\bar{x}}$ =

(1+ε)x. But, because *(x, y)* is not an MPSS, ray average product cannot remain constant as the input scale is slightly increased. Hence, ray average product must fall as the input scale is slight increased from *x*. Thus, locally diminishing returns to scale holds at every *(x, y)* $\in G$, when *x* is larger than the largest MPSS.

Identifying the Nature of Local Returns to Scale

Banker's Primal Approach

Banker (1984) developed the following important theorem that serves as a basis for identifying the nature of local returns to scale at the input-output bundle (x^0, y^0) if it is on the VRS frontier and at its efficient projection if it is an interior point[6].

Theorem 1: An input-output bundle (x^0, y^0) is an MPSS if and only if the optimal value of the objective function of a CCR-DEA model equals unity for this input-output combination.

Proof: Consider the input-oriented CCR DEA problem:

$$\begin{aligned}
&\min \theta \\
&s.t. \sum_{j=1}^{N} \lambda_j x^j \le \theta x^0; \\
&\quad \sum_{j=1}^{N} \lambda_j y^j \ge y^0; \qquad\qquad (30) \\
&\lambda_j \ge 0, (j = 1, 2, ..., N); \theta \; unrestricted.
\end{aligned}$$

A complete proof of this theorem requires us to show that (a) the optimal value θ^* must be unity if (x^0, y^0) *is* an MPSS and (b) θ^* cannot be unity if (x^0, y^0) *is not* an MPSS.

First, assume that (x^0, y^0) is an MPSS but θ^*<1 in (30), where the optimal solution is $(\theta^*, \lambda_j^*; j = 1, 2, ..., N).$ Then feasibility of the solution implies that

[6] See also Banker and Thrall (1992) and Banker et al (2004)

$$\sum_{j=1}^{N}\lambda_j^* x^j \le \theta^* x^0;$$
$$\sum_{j=1}^{N}\lambda_j^* y^j \ge y^0; \qquad (31)$$
$$\lambda_j^* \ge 0, (j=1,2,...,N).$$

Now define $k=\sum_{j=1}^{N}\lambda_j^*, \beta=\frac{\theta^*}{k}, \alpha=\frac{1}{k}$, and $\mu_j=\frac{\lambda_j^*}{k}$. Clearly, $\sum_{j=1}^{N}\mu_j=1$ and $\mu_j \ge 0, (j=1,2,...,N)$.

Thus,

$$\sum_{j=1}^{N}\mu_j x^j \le \beta x^0;$$
$$\sum_{j=1}^{N}\mu_j y^j \ge \alpha y^0; \qquad (32)$$
$$\sum_{j=1}^{N}\mu_j = 1;$$
$$\alpha,\beta > 0; \mu_j \ge 0, (j=1,2,...,N).$$

Therefore, $(\beta x^0, \alpha y^0)$ is a feasible input-output bundle under the VRS assumption. Further, $\frac{\alpha}{\beta}=\frac{1}{\theta^*}>1$ because θ^* has been assumed to be less than 1. This contradicts the assumption that (x^0, y^0) is an MPSS. Hence, part (a) is proven by contradiction.

Next suppose that (x^0, y^0) *is not* an MPSS but θ^* is equal to 1. Because (x^0, y^0) *is not* an MPSS there exist $\alpha, \beta,$ and μ_j $(j=1,2,...,N)$ such that

$$\sum_{j=1}^{N}\mu_j x^j \le \beta x^0;$$
$$\sum_{j=1}^{N}\mu_j y^j \ge \alpha y^0; \qquad (33)$$
$$\sum_{j=1}^{N}\mu_j = 1; \frac{\alpha}{\beta} > 1;$$
$$\alpha,\beta > 0; \mu_j \ge 0, (j=1,2,...,N).$$

Define $\lambda_j=\frac{\mu_j}{\alpha} (j=1,2,...,N)$. Then

$$\sum_{j=1}^{N} \lambda_j x^j \le \tfrac{\beta}{\alpha} x^0;$$
$$\sum_{j=1}^{N} \lambda_j y^j \ge y^0; \qquad (34)$$
$$\lambda_j \ge 0, (j = 1, 2, ..., N).$$

This shows that $\theta = \frac{\beta}{\alpha}$ is a feasible value of the objective function in the minimization problem in (30). But $\frac{\alpha}{\beta} > 1 \Rightarrow \frac{\beta}{\alpha} < 1$. This proves that the optimal value in (30) cannot be $\theta^* = 1$ unless (x^0, y^0) is an MPSS. This completes the proof of the theorem.

This theorem only determines whether (x^0, y^0) is an MPSS or not. It does not say anything directly about the nature of local returns to scale when it is not an MPSS. However, three important corollaries follow from the theorem:

1. If $k = \sum_{j=1}^{N} \lambda_j^* = 1,$ (x^0, y^0) is an MPSS and CRS holds locally.
2. If $k = \sum_{j=1}^{N} \lambda_j^* < 1,$ IRS holds locally at (x^0, y^0) or at its input-oriented efficient projection on to the VRS frontier if it is technically inefficient.
3. If $k = \sum_{j=1}^{N} \lambda_j^* > 1,$ DRS holds locally at (x^0, y^0) or at its input-oriented efficient projection on to the VRS frontier if it is technically inefficient.

The intuition behind these corollaries is quite simple. When *k* = 1, the optimal solution from the CRS problem in (30) is an optimal solution for the corresponding VRS problem. Because the CRS and VRS technical efficiency measures are identical, scale efficiency equals unity and (x^0, y^0) is an MPSS. Moreover, by virtue of part (a) of the theorem, θ^* equals unity and (x^0, y^0) is on the frontier. If $k \ne 1,$ the CRS input-oriented projection $(\theta^* x^0, y^0)$ is not a feasible solution for the corresponding VRS problem. But $\frac{1}{k}(\theta x^0, y^0)$ is both on the CRS and the VRS frontier. If *k* < 1, the input-oriented projection is to be scaled up to attain an MPSS and it lies in the IRS region. On the other hand,

if *k* > 1, it is scaled down to the MPSS and the input-oriented projection falls in the DRS region on the VRS frontier.

A potential problem with this method of returns to scale characterization is that there may be multiple optimal solutions to the DEA problem in (30) with the sum of λs greater than 1 in some and less than 1 in others. In that situation conflicting conclusions would be drawn depending on which optimal solution was obtained. This requires a modification of corollaries (2) and (3) as follows:

(2a) Locally increasing returns to scale holds if $k = \sum_{j=1}^{N} \lambda_j^* < 1$ at all optimal solutions of the CRS DEA problem in (30).

(3a) Locally diminishing returns holds if $k = \sum_{j=1}^{N} \lambda_j^* > 1$at all optimal solutions of the CRS DEA problem in (30).

This can be implemented in two steps. In step 1, the DEA problem in (30) is solved and the optimal value θ^* is determined. For (2a) above, in step 2 the following problem is solved:

$$\begin{aligned}
&\max \sum_{j=1}^{N} \lambda_j \\
&s.t. \sum_{j=1}^{N} \lambda_j x^j \le \theta^* x^0; \\
&\qquad \sum_{j=1}^{N} \lambda_j x^j \ge y^0; \\
&\qquad \lambda_j \ge 0, (j = 1, 2, ..., N).
\end{aligned} \tag{35}$$

If the maximum value of the objective function is less than 1, it can be concluded that $k = \sum_{j=1}^{N} \lambda_j^* < 1$at all optimal solutions of (30) . Similarly, in order to check for (3a) one minimizes the sum of λs in (31) and if the minimum is greater than 1, one can conclude that DRS holds locally.

A Dual Approach

Banker, Charnes, and Cooper (BCC) (1984) offer an alternative method of identifying local returns to scale from the following dual of the input-oriented VRS DEA problem:

$$\begin{aligned} &\max \quad v^{0\prime} y^0 - \delta_0 \\ &s.t. \quad v^{0\prime} y^j - \delta_0 \le u^{0\prime} x^j ,\ (j = 1, 2, ..., N); \\ &\qquad u^{0\prime} x^0 = 1; \\ &\qquad u^0 , v^0 \ge 0; \delta_0 \ unrestricted. \end{aligned} \qquad (36)$$

BCC have shown that

(i) CRS holds at (x^0, y^0) if at the optimal solution of (32) δ_0 is zero;

(ii) IRS holds at (x^0, y^0) if at the optimal solution of (32) δ_0 is < 0;

(iii) DRS holds at (x^0, y^0) if at the optimal solution of (32) δ_0 is > 0.

As in the case of Banker's approach, multiple optimal solutions pose a problem and the conditions (ii) and (iii) have to be appropriately modified.

A Nesting Approach

Färe, Grosskopf, and Lovell (FGL) (1985) consider a technology that lies in between CRS and the VRS technologies. They call it a non-increasing returns to scale (NIRS) technology. Under the assumption of NIRS

$$(x^0, y^0) \in T \Rightarrow (kx^0, ky^0) \in T \text{ for any } k \in (0, 1).$$

The DEA estimate of an NIRS production possibility set is

$$S^{NIRS} = \left\{ (x, y) : x \ge \sum_1^N \lambda_j x^j ; y \le \sum_1^N \lambda_j y^j ; \sum_1^N \lambda_j \le 1; \lambda_j \ge 0; j = 1, 2, ..., N \right\}. \qquad (37)$$

It may be noted that the frontiers of the CRS and NIRS production possibility sets coincide in the region of IRS. Similarly, the VRS and NIRS frontiers are identical in the DRS region. Therefore, when IRS holds at (x^0,y^0), in an input-oriented model $\theta_*^C = \theta_*^{NIRS} < \theta_*^V$ where the superscripts C, N, and V refer to CRS, NIRS, and VRS. Similarly, $\theta_*^C < \theta_*^{NIRS} = \theta_*^V$ implies DRS. Of course, in the case of CRS, all three estimates of technical efficiency equal unity.

Identifying Returns to Scale for inefficient unit

The concept of returns to scale is meaningful only when the relevant input-output bundle lies on the frontier of the production possibility set. For an inefficient bundle, one must consider its efficient projection – either input- or output-oriented. Unless similar returns to scale are found at both projections, one cannot conclusively determine the returns to scale at the observed input-output bundle.

The following DEA problem considered by Cooper, Thompson, and Thrall (1996) can be used not only to determine whether an input-output bundle *(x^0, y^0)* is an MPSS but also to identify the bundle *(x^*, y^*)* which is an MPSS for *(x^0, y^0)* :

Maximize $\rho = \frac{\alpha}{\beta}$

Subject to

$$\sum_j \lambda_j x^j \le \beta x^0;$$

$$\sum_j \lambda_j y^j \ge \alpha y^0; \qquad (38)$$

$$\sum_j \lambda_j = 1;$$

$$\alpha, \beta, \lambda_j \ (j = 1,2,...,N) \ge 0.$$

Because *(x^0, y^0)* is assumed to be a feasible input-output bundle, $(\alpha = \beta = \rho = 1)$ is a feasible solution for this problem. Hence, the optimal value ρ^* is always greater than or equal to 1.When $\rho^* = \frac{\alpha^*}{\beta^*}$ exceeds unity, we know that *(x^0, y^0) is not an MPSS*. But we can also conclude that $(\beta^* x^0, \alpha^* y^0)$ *is an MPSS*

As such, the objective function is nonlinear. However, it can be easily transformed into a linear programming problem. Define $t = \frac{1}{\beta}$ and $\mu_j = t\lambda_j$ (*j = 1, 2,…,N*). Note that non-negativity of β and λ_js ensures that *t* and μ_js are also non-negative. Problem (5) can, therefore, be reformulated as the following linear programming problem:

$$\begin{aligned} &\text{Maximize } \rho \\ \text{subject to } &\sum_j \mu_j x^j \le x^0; \\ &\sum_j \mu_j y^j \ge \rho y^0; \qquad (39) \\ &\sum_j \mu_j = t; \\ &t, \mu_j \; (j = 1,2,...,N) \ge 0. \end{aligned}$$

From the optimal solution of this problem we can derive $\beta^* = \frac{1}{t^*}$ and $\alpha^* = \frac{\rho^*}{t^*}$. One can then infer the nature of returns to scale from these values of α^* and β^*. It may be pointed out here that because the only restriction on *t* is non-negativity, (27) is simply the output-oriented CCR DEA problem and $\frac{1}{\rho^*}$ is the same as the output-oriented CRS technical efficiency $\tau_y^C(x^0, y^0)$.

When the bundle (x^0, y^0) is not itself an MPSS, $\rho^* > 1$so that $\alpha^* > \beta^*$. If the MPSS is unique, there are five different possibilities: (i) $1 < \beta^* < \alpha^*$; (ii) $\beta^* < \alpha^* < 1$; (iii) $\beta^* = 1 < \alpha^*$; (iv) $1 < \beta^* = \alpha^*$, and (v) $1 < \beta^* < \alpha^*$. When the MPSS is unique, both input- and output-oriented projections of the bundle (x^0, y^0) fall in the region of IRS. In this case, the unit is conclusively too small relative to its MPSS. Similarly, if $\beta^* < \alpha^* < 1$, both input- and output-oriented projections fall in the region of DRS. The implication is that the unit is too large. When $\beta^* = 1 < \alpha^*$, the input scale corresponds to the MPSS but the output scale is too small. The opposite is true when $1 < \beta^* = \alpha^*$. Finally, in the intermediate case, where $\beta^* < 1 < \alpha^*$, the input scale is bigger than the MPSS and the output-oriented projection falls in the region of DRS. At the same time, the input-scale is smaller than the MPSS and the input-oriented projection falls in the region of IRS.

Zhu (2001) uses a single-input single-output example to partition the interior of the production possibility set into six different regions for returns to scale classification of inefficient production units[7]. In three out of these six regions both input- and output-oriented efficient projections exhibit the same returns to scale: increasing, constant, or diminishing. In the remaining three, increasing returns at the input-oriented projection combines with constant or diminishing returns at the output-oriented projection, or constant returns at the input-oriented projection is associated

[7] See also the earlier paper by Seiford and Zhu (1999).

with diminishing returns at the output-oriented. In order to correctly locate an inefficient unit in the appropriate region, one has to ascertain returns to scale at both projections.

The Case of Multiple MPSS

Next we consider the possibility of multiple MPSS. This is depicted graphically in Figure 3. Here both *C1* and *C2* are MPSS and so are their convex combinations lying on the line segment connecting them. At *C1*, (α_1^*, β_1^*) is the smallest MPSS. Similarly, (α_2^*, β_2^*) at *C2* is the largest MPSS. It is obvious that when (39) has a unique optimal solution (in particular, t^* is unique), there cannot be multiple MPSS. For multiple optimal solutions, the largest $t^* = \sum_j \mu_j^*$ across all optimal solutions of (39) corresponds to the smallest MPSS, β_1^* . Similarly, β_2^* corresponds to the smallest $t^* = \sum_j \mu_j^*$ at an optimal solution.

Note that across all optimal solutions the value of the objective function is the same *(ρ^*)*. Hence, $\beta_1^* = \frac{1}{t_1^*}$, where

$$t_1^* = \max \sum_j \mu_j$$

s.t.

$$\begin{aligned} &\sum_j \mu_j x^j \le x^0; \\ &\sum_j \mu_j y^j \ge \rho^* y^0; \qquad (40) \\ &\mu_j \ (j = 1,2,\ldots,N) \ge 0. \end{aligned}$$

Similarly, $\beta_2^* = \frac{1}{t_2^*}$, where

$$t_2^* = \min \sum_j \mu_j$$

s.t.

$$\begin{aligned} &\sum_j \mu_j x^j \le x^0; \\ &\sum_j \mu_j y^j \ge \rho^* y^0; \qquad (41) \\ &\mu_j \ (j = 1,2,\ldots,N) \ge 0. \end{aligned}$$

Once β_1^* and β_2^* have been determined from (8) and (9), the corresponding values of α are readily obtained as $\alpha_1^* = \rho^* \beta_1^*$ and $\alpha_2^* = \rho^* \beta_2^*$.

As shown in Figure 3, the set of output-input scales *(α, β)* for which the input-output bundles *($\beta x^0, \alpha y^0$)* are feasible can be partitioned into six different regions defined below:

(i) In region (1) towards the southwest of the smallest MPSS (*C_1*), $(\beta < \beta_1^*; \alpha < \alpha_1^*)$. When *($x^0$, y^0)* falls in this region, $1 < \beta_1^* < \alpha_1^*$. Hence, increasing returns to scale holds unambiguously.

(ii) In region (2) to the northeast of the largest MPSS (*C_2*), $(\beta_2^* < \beta; \alpha_2^* < \alpha)$. If *($x^0$, y^0)* falls in this region, $\beta_1^* < \alpha_1^* < 1$. Diminishing returns to scale holds unambiguously in this region.

(iii) In region (3), $\beta_1^* < \beta < \beta_2^*$ while $\alpha_1^* < \alpha < \alpha_2^*$. Points in this region lie between the smallest and the largest MPSS. It is interesting to note, that even if the point *(α*= 1, *β*= 1*)* is not technically efficient and lies below the *C_1C_2* line, both the input- and the output-oriented projection of the inefficient bundle will fall in the region of constant returns to scale. Thus, there is no scale inefficiency in this region even though there may be technical inefficiency,.

(iv) In region (4), $\beta_2^* < \beta; \alpha < \alpha_1^*$. When the actual input-output bundle lies here, $\beta_2^* < 1 < \alpha_1^*$. The input bundle *$x^0$* is larger than the largest MPSS hence the output oriented projection falls in the area of diminishing returns. At the same time, the actual output bundle is smaller than the smallest MPSS. Hence, increasing returns to scale holds at the input oriented projection. Thus, returns to scale cannot be unambiguously defined at the actual input-output bundle.

(v) In region (5a), $\beta_1^* < \beta < \beta_2^*$ but $\alpha < \alpha_1^*$. When the actual input-output bundle lies here, *y^0* is smaller than the smallest MPSS and the input oriented projection falls in the area of increasing returns. At the same time, the actual input bundle lies between the smallest and the largest MPSS. Hence, constant returns to scale holds at the output oriented projection. Here also the returns to scale characterization depends on the orientation.

(vi) In region (5b), $\beta_2^* < \beta$ while $\alpha_1^* < \alpha < \alpha_2^*$. When the actual input-output bundle lies here, x^0 is larger than the largest MPSS. Hence the output oriented projection falls in the area of diminishing returns. At the same time, the actual output bundle lies between the smallest and the largest MPSS. Hence, constant returns to scale holds at the input oriented projection. Here the input bundle is too large. But the actual output bundle, if produced from the technically efficient input bundle would correspond to an MPSS..

Output- or Input-oriented?

Except in the case of globally constant returns to scale, output- and input-oriented technical efficiency measures would differ for the same firm. An important question is how to decide which measure is preferable. As a general rule, the answer depends on whether output augmentation is more important that input conservation in a specific context. In many situations, however, there is no clear cut prority. A rule of thumb would then be to select the orientation that yields a *lower measure of efficiency* under the VRS assumption. The logic behind this criterion is the fact the corresponding efficient projection would have a higher level of scale efficiency. This can be explained by a simple 1-input 1-output example. Consider a technically inefficient input-output combination (x_0, y_0). Now suppose that the output-oriented efficient projection is $(x_0, \varphi^* y_o)$ while in input-oriented projection is $(\theta^* x_0, y_o)$. Thus the corresponding technical efficiency measures are $\tau_y = \frac{1}{\varphi^*}$ and $\tau_x = \theta^*$. Assume arbitrarily that $\tau_y < \tau_x$. This implies $\frac{1}{\varphi^*} < \theta^*$ or $\frac{1}{\theta^*} < \varphi^*$. Therefore, $\frac{y_0}{\theta^* x_0} < \frac{\varphi^* y_0}{x_0}$. This shows that average productivity is higher at the output-oriented efficient projection than at the input-oriented projection of (x_0, y_0).

4. Non-Radial Measures of Technical Efficiency

In an output-oriented analysis of technical efficiency the objective is to produce the maximum output from a given quantity of inputs. For this we first define the *(producible) output set* of any given input bundle. For the input bundle x^0 the output set

$$P(x^0) = \{ y: (x^0, y) \in T\} \qquad (42)$$

consists of all output bundles that can be produced from x^0.

Because there are different output sets for different input bundles, the production possibility set is equivalently characterized by a family of output sets. Each output set is a subset of the *m*-

dimensional output space. The following properties of output sets follow from the relevant assumptions made about the production possibility set.

(P1) If (x^j, y^j) is an actually observed input-output combination, then $y^j \in P(x^j)$.

(P2) If $y^0 \in P(x^0)$ and if $x^1 \geq x^0$, then $y^0 \in P(x^1)$.

(P3) If $y^0 \in P(x^0)$ and if $y^1 \leq y^0$, then $y^1 \in P(x^0)$.

(P4) Each output set $P(x)$ is convex.

The *output isoquant* of any input bundle x^0 can be defined as

$$\overline{P}(x^0) = \{y : y \in P(x^0) \text{ and } \lambda y \notin P(x^0) \text{ if } \lambda > 1\}. \qquad (43)$$

Thus, if $y^0 \in \overline{P}(x^0)$, then the output-oriented radial technical efficiency of the pair (x^0, y^0) equals unity because it is not possible to increase *all* outputs holding the input bundle unchanged. This does not, of course, rule out the possibility that some individual components of the y^0 output bundle can be increased.

The *efficient subset* of the output isoquant of x^0, on the other hand, is

$$P^*(x^0) = \{y : y \in P(x^0) \text{ and } y' \notin P(x^0) \text{ if } y' \geq y^0\}. \qquad (44)$$

Thus, an output-oriented radial technically efficient projection of y^0 produced from x^0 onto $P(x^0)$ may include slacks in individual outputs. But no such slacks may exist if the projection is onto $P^*(x^0)$.The radial measure of output-oriented technical efficiency does not reflect any unutilized potential for increasing individual outputs. Again, as is shown below a non-radial output-oriented measure does take account of all potential increase in any component of the output bundle.

The problem of slacks in any optimal solution of a radial DEA model arises because we seek to expand all outputs or contract all inputs by the same proportion. In non-radial models, one allows the individual outputs to increase or the inputs to decrease at different rates. Färe and Lovell (1978) introduced the following input-oriented, *non-radial* measure of technical efficiency called the Russell measure:

$$\rho_x(x^0, y^0) = \min \tfrac{1}{n} \sum_i \theta_i$$

$$\text{s.t.} \quad \sum_j \lambda_j y_{rj} \geq y_{r0}; \ (r = 1,2,\ldots,m); \qquad (45)$$

$$\sum_j \lambda_j x_{ij} \leq \theta_i x_{i0}; \quad \textit{(i= 1, 2,…, n);}$$

$$\sum_j \lambda_j = 1; \quad \lambda_j \geq 0; \quad \textit{(j =1, 2,…, N).}$$

When input slacks do exist at the optimal solution of a radial DEA model, the non-radial Russell measure in (39) falls below the conventional measure obtained from an input-oriented BCC model (6). Because the radial projection is always a feasible solution for (8), $\rho_x \leq \tau_x$. That is, the non-radial Russell measure of technical efficiency never exceeds the corresponding radial measure.

The analogous output-oriented non-radial VRS measure of technical efficiency is:

$$RM_y(x^0, y^0) = \frac{1}{\rho_y},$$

where $\rho_y = \max \tfrac{1}{m} \sum_r \phi_r$

$$\text{s.t.} \quad \sum_j \lambda_j y_{rj} \geq \phi_r y_{r0}; \quad \textit{(r = 1,2,…,m);}$$

$$\sum_j \lambda_j x_{ij} \leq x_{i0}; \quad \textit{(I = 1, 2,…, n);} \qquad (46)$$

$$\sum_j \lambda_j = 1; \quad \lambda_j \geq 0; \quad \textit{(j =1, 2,…, N).}$$

While no input slacks can exist at the optimal solution of (39), presence of any output slack is not ruled out. Similarly, input slacks may remain at the optimal solution of (40).

Graph Efficiency Measures

All of the technical efficiency measures considered above are either output- or input-oriented. Instead of focusing exclusively on increasing outputs or reducing inputs, one may wish to achieve both objectives simultaneously. A problem with a graph efficiency measure is that the benchmark input-output bundle selected on the frontier depends on the relative importance attached to input reduction vis-à-vis output expansion. In the extreme case of input-orientation, output expansion is given zero weight. Conversely, 100% weight assigned to output expansion leads to the output-oriented projection. As discussed before, the relative importance of outputs and inputs is usually a matter of judgment by the analyst. Färe, Grosskopf, and Lovell (1985, 1994) introduced the so called

Graph Hyperbolic measure of efficiency by selecting a benchmark on the frontier where is actual output bundle (y^0) is scaled up as δy^0 while the observed input bundle (x^0) is scaled down as $(\frac{1}{\vartheta}x^0)$. It is easy to see that in a single output single input case bothe the actual input-output bundle (x^0, y^0) and its efficient projection $(\frac{1}{\delta}x^0, \delta y^0)$ will lie on a rectangular hyperbola. This explains the name *Graph Hyperbolic* efficiency.

The VRS DEA LP problem for measuring the graph efficiency is

$$\begin{aligned}
&\max\ \delta \\
&s.t. \sum_{j=1}^{N}\lambda_j y^j \geq \delta y^0; \\
&\quad \sum_{j=1}^{N}\lambda_j x^j \leq \tfrac{1}{\delta}x^0; \\
&\quad \sum_{j=1}^{N}\lambda_j = 1; \qquad\qquad (47) \\
&\quad \lambda_j \geq 0, (j = 1, 2, ..., N); \\
&\quad \delta\, unrestricted.
\end{aligned}$$

A problem with the DEA problem (47) is that it is a non-linear programming problem. It is possible, however, to use a linear approximation for $\frac{1}{\delta}$ in the input constraint. Note that a first order Taylor's series approximation to $f(\delta) = \frac{1}{\delta}$ at the point of approximation δ_0 is

$$f(\delta) \simeq f(\delta_0) + (\delta - \delta_0) f'(\delta_0) = \tfrac{1}{\delta_0} - \tfrac{1}{\delta_0^2}(\delta - \delta_0).$$

Using $\delta_0 = 1$ as the point of approximation, $\frac{1}{\delta} \approx 2 - \delta$ and the problem in (41) can be linearized as

$$\begin{aligned}
&\max\ \delta \\
&s.t.\ \sum_{j=1}^{N}\lambda_j y^j \ge \delta y^0; \\
&\quad \sum_{j=1}^{N}\lambda_j x^j + \delta x^0 \le 2x^0; \\
&\quad \sum_{j=1}^{N}\lambda_j = 1; \qquad\qquad (48) \\
&\quad \lambda_j \ge 0, (j = 1, 2, ..., N); \\
&\quad \delta\, unrestricted.
\end{aligned}$$

It should be noted that if (x^0, y^0) is quite far from the frontier, $\delta_0 = 1$may lead to a very poor approximation. This is likely to true if either the input- or the output-oriented efficiency is low. In that case, the linear approximation may need to be applied iteratively with the optimal value of δ providing the point of approximation for each successive iteration.

In the case of CRS, the optimization problem in (47) can be handled quite easily. Without the restriction $\sum_{j=1}^{N}\lambda_j = 1,$ (47) can be written as

$$\begin{aligned}
&\max\ \delta^2 \\
&s.t.\ \sum_{j=1}^{N}\delta\lambda_j y^j \ge \delta^2 y^0; \\
&\quad \sum_{j=1}^{N}\delta\lambda_j x^j \le x^0; \\
&\qquad\qquad\qquad\qquad (49) \\
&\quad \lambda_j \ge 0, (j = 1, 2, ..., N); \\
&\quad \delta\, unrestricted.
\end{aligned}$$

Defining $\varphi = \delta^2$ and $\mu_j = \delta\lambda_j,$ the DEA problem in (43) becomes the standard CCR output oriented problem

$$
\begin{aligned}
&\max\ \varphi \\
&s.t. \sum_{j=1}^{N} \mu_j y^j \geq \varphi y^0; \\
&\quad \sum_{j=1}^{N} \mu_j x^j \leq x^0; \qquad\qquad (50) \\
&\quad \mu_j \geq 0, (j = 1, 2, ..., N); \\
&\quad \varphi \ \ unrestricted.
\end{aligned}
$$

Note that φ should be constrained to be strictly positive. But given that $\varphi = 1$ is always a feasible solution, it would be a non-binding constraint. The optimal value of δ can be computed as $\sqrt{\varphi}$.

Pareto-Koopmans Efficiency

An input-output combination (x^0, y^0) is not Pareto-Koopmans efficient if it violates either of the following efficiency postulates:

(i) It is not possible to increase any output in the bundle y^0 without reducing any other output and/or without increasing any input in the bundle x^0; or

(ii) It is not possible to reduce at least any input in the bundle x^0 without increasing any other input and/or without reducing any output in the bundle y^0.

Clearly, unless $RM_x(x^0, y^0) = RM_y(x^0, y^0) = 1$, at least one of the two efficiency postulates is violated and (x^0, y^0) is not Pareto-Koopmans efficient. Input-output bundle (x^0, y^0) is Pareto-Koopmans efficient, when both of the following conditions hold:

$$\text{(i) } x^0 \in V^*(y^0) \qquad \text{and} \qquad \text{(ii) } y^0 \in P^*(x^0).$$

Thus, non-radial technical efficiency (whether input-oriented or output-oriented) by itself does not ensure overall Pareto efficiency.

A non-radial Pareto-Koopmans measure of technical efficiency of the input-output pair *(x^0, y^0)* can be computed as:

$$\gamma(x^0, y^0) = \min \frac{\frac{1}{n}\sum_i \theta_i}{\frac{1}{m}\sum_r \phi_r}$$

$$\text{s.t.} \qquad \sum_{j=1}^{N} \lambda_j y_{rj} \geq \phi_r y_{r0}; \qquad (r = 1,2,..., m);$$

$$\sum_{j=1}^{N} \lambda_j x_{ij} \leq \theta_i x_{i0}; \qquad (i = 1,2,...,n); \qquad (51)$$

$$\sum_{j=1}^{N} \lambda_j = 1; \qquad \lambda_j \geq 0; \quad (j = 1,2,...,N).$$

Note that the efficient input-output projection *(x^*, y^*)* satisfies

$$x^* = \sum_{j=1}^{N} \lambda_j^* x^j \leq x^0 \qquad \text{and} \qquad y^* = \sum_{j=1}^{N} \lambda_j^* y^j \geq y^0 \ .$$

Thus, *(x^0, y^0)* is Pareto-Koopmans efficient, if and only if $\phi_r^* = 1$ for each output *r* and $\theta_i^* = 1$ for each input *i*, implying $\gamma(x^0, y^0) = 1$.We can visualize the Pareto-Koopmans global efficiency measure as the product of two factors. The first is the input-oriented component

$$\gamma_x = \tfrac{1}{n} \sum_i \theta_i \qquad (52)$$

and the second is an output-oriented component

$$\gamma_y = \frac{1}{\frac{1}{m} \sum_r \phi_r} . \qquad (53)$$

Thus,

$$\gamma(x^0, y^0) = \gamma_x . \gamma_y . \qquad (54)$$

A Slack Based Measure of Efficiency

Tone (1997) introduced essentially the same measure of overall efficiency and called it a slack based measure (SBM)[8]. Tone's SBM is

[8] This is the same as the *extended Russell Measure* of Pastor, Ruiz, and Sirvent (1999).

$$\rho = \min \frac{1-\frac{1}{n}\sum_{i=1}^{n}\frac{s_i^-}{x_{i0}}}{1-\frac{1}{m}\sum_{r=1}^{m}\frac{s_r^+}{y_{r0}}}$$

$$s.t. \sum_{j=1}^{N}\lambda_j x_{ij} + s_i^- = x_{i0}, (i = 1,2,...,n);$$

$$\sum_{j=1}^{N}\lambda_j y_{rj} + s_r^+ = y_{r0}, (r = 1,2,...,m); \qquad (55)$$

$$\sum_{j=1}^{N}\lambda_j = 1;$$

$$\lambda_j, s_i^-, s_r^+ \geq 0, (j = 1,2,...,N; i = 1,2...,n; r = 1,2,...,m).$$

In this formulation, s_i^- is the *total* slack (rather than the *radial* slack) in input *i*. Similarly, s_r^+ is the *total* slack in output *r*. Now consider the benchmark bundle (x^*, y^*) where $x_i^* = \theta_i x_{i0} = x_i^0 - s_i^-$ for input *i* and $y_r^* = \varphi_r y_{r0} = y_r^0 + s_r^+$ for output *r*. It is obvious that $\theta_i = 1 - \frac{s_i^-}{x_{i0}}$ and $\varphi_r = 1 + \frac{s_r^+}{y_{r0}}$. That is, the objective function in (51) is the same as that in (55). Similarly, the constraints are the same in both problems except that non-negativity of the slacks implies that in the SBM, the *θ*s cannot exceed unity and the *φ*s cannot be lower than 1. No such constraints are implicit in (51).

Linearization of the Pareto-Koopmans DEA Problem

The objective function in (51) is non-linear. Tone transformed this linear fractional functional programming problem into an LP problem by normalizing the denominator to unity. Alternatively, as shown in Ray (2004), one may replace the objective function by a linear approximation.

Define

$$\gamma(x^o, y^o) = f(\theta, \varphi) \qquad (56)$$

Using $\theta_i^0 = 1$ for all *i* and $\phi_r^0 = 1$ for all *r* as the point of approximation,

$$f(\theta, \varphi) \approx 1 + \tfrac{1}{n}\sum_i \theta_i - \tfrac{1}{m}\sum_r \phi_r . \qquad (57)$$

We may, therefore, replace the objective function in (51) by (57) and solve (51) iteratively using the optimal solution from each iteration as the point of approximation for the next iteration until convergence. Once we obtain the optimal (θ^*, ϕ^*) from this problem, we evaluate

$$\gamma(x^0, y^0) = \frac{\frac{1}{n}\sum_i \theta_i^*}{\frac{1}{m}\sum_r \phi^*_r} \qquad (58)$$

as a measure of the Pareto-Koopmans efficiency[9] of *(x^0, y^0).*

Apart from an overall measure, (51) also provides information about the potential for reducing individual inputs (θ_i^*) and increasing individual outputs (φ_r^*). Also a decomposition of (51) into the input- and output-oriented components can be obtained from (12[10]).

5. Distance Functions

Shephard (1953) defined the *Distance Function* evaluated at any non-negative input-output bundle (*x*, *y*) as

$$D(x, y) = \min \beta : (x, \tfrac{1}{\beta} y) \in T. \qquad (59)$$

It may be noted that the bundle (*x*, *y*) itself may not feasible and may lie outside the production possibility set. Shephard assumed only that every input-output bundle can be projected on to the frontier of the production possibility set by appropriately scaling (up or down) the input or the output bundle. This was described as the *attainability* postulate.

If an output bundle *y* cannot be produced from the input bundle *x*, then the attainability assumption ensures that any appropriate scaling down (without altering the mix of outputs) would result in a feasible bundle. That would imply a value of β greater than 1. On the other hand, if (*x*, *y*) is in the interior of the production possibility set, the output bundle can be scaled up and still be producible from the input bundle *x*. In that case, β would be less than 1.

Two things emerge out of the above. First, an alternative characterization of the production possibility set is

$$T = \{(x, y) : D(x, y) \le 1\}. \qquad (60)$$

Second, β in (59) is simply the inverse of φ in (8). Thus, the Shephard Distance Function is the same as the Farrell output-oriented technical efficiency, τ_y.

[9] For empirical applications of this Pareto Koopmans measure see Ray and Jeon (2009) and Ray and Ghosh (2014).

[10] See Ray (2004) appendix to Chapter 2 for a proof of these properties.

The Distance Function defined in (59) is more accurately described as the *output Distance Function*

$$D^O(x,y) = \min \beta : (x, \tfrac{1}{\beta} y) \in T. \qquad (61)$$

The *input Distance Function* can, analogously, be defined as

$$D^I(x,y) = \max \mu : (\tfrac{1}{\mu} x, y) \in T. \qquad (62)$$

It is clear that D^I *(x, y)* is the inverse of the input oriented Farrell efficiency, τ_x. Moreover, under constant returns to scale, the *output* and *input* Distance Functions are inverses of each other. The following properties of the Distance Functions should be noted.

The *output* Distance Function, $D^O(x,y)$ is

(a) homogeneous of degree 1 in *y*;

(b) increasing (non-decreasing) in *y*;

(c) decreasing (non-increasing) in *x*;

(d) convex in *y*.

Similarly, the *input* Distance Function, $D^I(x,y)$, is

(a) homogeneous of degree 1 in *x*;

(b) increasing (non-decreasing) in *x*;

(c) increasing (non-decreasing) in *y*;

(d) convex in *x*.

<u>Directional Distance Function</u>

Chambers, Chung, and Färe (CCF) (1996) introduced the Nerlove-Luenberger *Directional Distance Function*:

$$\vec{D}(x,y;g^x,g^y) = \max \beta : (x + \beta g^x, y + \beta g^y) \in T. \qquad (63)$$

Here, (g^x, g^y) is an arbitrary point that serves to define the direction along which the input-output bundle (*x*, *y*) is projected on to the frontier[11]. It is important to recognize that (g^x, g^y) need not be a feasible input-output bundle (or for that matter, even non-negative!). Its only role is to define the direction along which the (*x*, *y*) bundle is to be projected on to the frontier. Of course, if (*x*, *y*) is already on the frontier, β will be equal to 0. If one selects $(g^x = -x, g^y = y)$, the Directional Distance Function becomes

[11] See also Färe, R. and S. Grosskopf (2000)

$$\vec{D}(x,y;-x,y)=\max\beta:\left((1-\beta)x,(1+\beta)y\right)\in T. \qquad (64)$$

In this case, β is the maximum proportionate reduction in all inputs simultaneously feasible with the same proportionate increase in all outputs. For this reason, it is sometimes described as the *proportional* Distance Function. Further, if one selects, $(gx=0,g^y=y)$, the Directional Distance Function coincides with the (inverse of) the output-oriented Shephard Distance Function. On the other hand, $(g^x=-x,g^y=0)$ leads to the input-oriented Shephard Distance Function.

The DEA LP problem for measuring the Directional Distance Function shown in (61) above is

$$\begin{aligned}
&\vec{D}(x^0,y^0;-x^0,y^0)=\max\beta\\
&s.t. \sum_{j=1}^{N}\lambda_j x^j+\beta x^0\le x^0;\\
&\sum_{j=1}^{N}\lambda_j y^j-\beta y^0\ge y^0; \qquad (65)\\
&\sum_{j=1}^{N}\lambda_j=1;\\
&\lambda_j\ge 0,(j=1,2,...,N);\beta\, unrestricted.
\end{aligned}$$

As elsewhere, the restriction $\sum_{j=1}^{N}\lambda_j=1$ is removed when CRS is assumed. It may be noted that even though β is unrestricted in principle, in practice it can never exceed 1 because the target input bundle (1-*β*)x^0 would otherwise become negative. Under the CRS assumption, the Directional Distance Function, β, can be easily derived from the output expansion factor, φ, in the CCR DEA problem. More specifically, $\varphi=\frac{1+\beta}{1-\beta}$. Alternatively, $\beta=\frac{\varphi-1}{\varphi+1}$. It should be noted that β is a measure of technical *inefficiency*. Also, like the CCR/BCC DEA measures, the CCF Directional Distance Function is also a radial measure because all inputs are scaled down by the factor (1-*β*) while all outputs are scaled up by the factor (1+*β*). Individual output- and/or input slacks may exist at the optimal solution of the DEA LP problem in (65). Hence, a value of β equal to 0 does not guarantee Pareto-Koopmans efficiency.

Geometric Distance Function

Portela and Thanassoulis (2004) introduced the Geometric Distance Function that provides a non-radial Pareto-Koopmans measure of technical efficiency. Consider the input bundle

$x^0 = (x_1^0, x_2^0, ..., x_n^0)$ and the output bundle $y^0 = (y_1^0, y_2^0, ..., y_m^0)$. The Geometric Distance Function can be evaluated as

$$\gamma(x^0, y^0) = \min \frac{\overset{n}{\underset{i=1}{\Pi}} (\theta_i)^{\alpha_i}}{\overset{m}{\underset{r=1}{\Pi}} (\varphi_r)^{\beta_i}}$$

$$\begin{aligned} s.t.\ &\sum_{j=1}^{N} \lambda_j x_i^j \le \theta_i x_i^0,\ (i = 1, 2, ..., n); \\ &\sum_{j=1}^{N} \lambda_j y_r^j \ge \varphi_r y_r^0,\ (r = 1, 2, ..., m); \\ &\sum_{j=1}^{N} \lambda_j = 1; \qquad\qquad (66) \\ &\theta_i \le 1;\ (i = 1, 2, ..., n); \\ &\varphi_r \ge 1, (r = 1, 2, ..., m); \\ &\lambda_j \ge 0, (j = 1, 2, ..., N). \end{aligned}$$

The exponents α_i and β_r are, respectively, the predetermined weights assigned to the individual inputs and outputs. The weights are non-negative and satisfy $\sum_i \alpha_i = \sum_r \beta_r = 1$. Thanassoulis and Portela set each α_i equal to $\frac{1}{n}$ and each β_r equal to $\frac{1}{m}$.

The only difference between the Geometric Distance Function in (63) above and the Pareto-Koopmans efficiency measure in (51) is that one is a ratio of the arithmetic means of the θs and the φs while the other is a ratio of their respective geometric means. Also, Portela and Thanassoulis (2004) restrict the θs to be no greater than and the φs to be no less than unity. Removing these restrictions would allow greater flexibility by allowing inputs to increase if that would permit and even greater increase in outputs or the outputs to decline if inputs decline even more.

As in the case of (51), the objective function in (63) also is nonlinear. However, taking its natural log one gets $\ln \gamma = \frac{1}{n}\sum_{i=1}^{n} \ln \theta_i - \frac{1}{m}\sum_{r=1}^{m} \ln \varphi_r$ which can be linearized at $(\theta_i = 1; i = 1, 2, ..., n; \varphi_r = 1; r = 1, 2, ..., m)$ as $\ln \gamma \approx \frac{1}{n}\sum_{i=1}^{n} \theta_i - \frac{1}{m}\sum_{r=1}^{m} \varphi_r$. One can set this up as a (approximate) linear objective function to iteratively solve the DEA optimization problem in (63).

6. Data Transformation

In practical applications, the same input or output quantities can, often, be measured in alternative units. The cultivated area may be measured in acres or in hectares. Oil may be measured in gallons or in liters. Output may be measured in lbs or kilograms. A change in the unit of measurement is essentially a change in *scale.* An efficiency measure is *scale invariant* if a change in scale does not alter the measured efficiency of the same input-output bundle[12].

The CCR and BCC technical efficiency measures (both input and output-oriented) are scale invariant. Consider the CCR output-oriented model first. Suppose that the scale of measurement of all (or some) of the individual inputs and outputs are changed. Specifically, the new measure of input *i* is $\tilde{x}_i = a_i x_i$. Similarly, output *r* is measured as $\tilde{y}_r = b_r y_r$.For the transformed data, the CCR output oriented DEA problem will be

$$\begin{aligned} &\max \varphi \\ &s.t. \sum_{j=1}^{N} \lambda_j \tilde{x}_i^j \geq \tilde{x}_i^0 (i = 1, 2, ..., n); \\ &\quad \sum_{j=1}^{N} \lambda_j \tilde{y}_r^j \geq \varphi \tilde{y}_r^0, (r = 1, 2, ..., m); \\ &\lambda_j \geq 0, (j = 1, 2, ..., N) \varphi \ \ unrestricted. \end{aligned} \tag{67}$$

But the input and output constraints actually are

$$\begin{aligned} &\sum_{j=1}^{N} \lambda_j (a_i x_i^j) \geq a_i x_i^0 (i = 1, 2, ..., n); \\ &\sum_{j=1}^{N} \lambda_j (b_r y_r^j) \geq \varphi b_r y_r^0, (r = 1, 2, ..., m). \end{aligned}$$

Hence, cancellation of common factors on both sides of the inequalities reduces the transformed DEA problem (64) to the original CCR output oriented problem (14). For the BCC output oriented problem there is the additional restriction that the λs add up to unity. But data transformation does not affect that constraint in any way. Hence the BCC output oriented technical efficiency measure is also scale invariant. Proof of scale invariance of the input oriented measures (both CCR and BCC) will be analogous.

[12] See Ali and Seiford (1990) and Lovell and Pastor (1995).

A different kind of transformation known as *translation* of the origin is one where some constant is added to (or subtracted from) any input or output quantity of all firms in the sample. Data translation is common in applications where some input or output values are found to be negative in the data set. A constant is added to all observations of that input or output to ensure non-negativity of the data.

We now show that the CCR DEA efficiency measures (whether input or output oriented) are not translation invariant. For this, suppose that the transformed input-output data are $\tilde{x}_i^j = x_i^j + c_i (i = 1, 2, ..., n)$ and $\tilde{y}_r^j = y_r^j + d_r (r = 1, 2, ..., m).$ The CCR DEA problem with the transformed data actually is

$$\begin{aligned} &\max \varphi \\ &s.t. \sum_{j=1}^{N} \lambda_j (x_i^j + c_i) \geq x_i^0 + c_i (i = 1, 2, ..., n); \\ &\quad \sum_{j=1}^{N} \lambda_j (y_r^j + d_r) \geq \varphi (y_r^0 + d_r), (r = 1, 2, ..., m); \\ &\lambda_j \geq 0, (j = 1, 2, ..., N) \varphi \ \ unrestricted. \end{aligned} \qquad (68)$$

It is obvious that the additional term $\sum_{j-1}^{N} \lambda_j c_i$ on the left hand side of the input constraint does not cancel with *c_i* on the right hand side. Nor does $\sum_{j-1}^{N} \lambda_j d_r$ cancel φd_r in the output constraints.

Hence, the problem in (68) will not have the same optimal solution as the original CCR problem in (14). In a similar manner, it can be shown that the CCR input oriented problem is also non translation invariant.

Next consider the BCC problems where VRS is captured by the additional constraint that the λs add up to unit. Given, $\sum_{j=1}^{N} \lambda_j = 1, \sum_{j=1}^{N} \lambda_j c_i = c_i.$ Thus, the input constraints in the output-oriented BCC problem with transformed data are the same as the input constraints in the corresponding problem before data translation. However, output constraints will differ unless each d_r equals 0. That is, there is no output translation. We conclude, therefore, the BCC output oriented DEA model is invariant to *input translation*. Similarly, the BCC input-oriented DEA problem is invariant to *output translation.*

7. Weak Disposability and Bad Output

In many cases production results in some *bad* or undesirables output side by side with the *good* or desirable output. In manufacturing, production of the desired output (like industrial machinery or steel) leaves the firm with some industrial waste potentially damaging to the environment. Generation of power at an electrical utility plant also results in emission of smoke and polluting particulates in the atmosphere. Traditionally, productivity and efficiency analysts have focused solely on the quantity of the good output produced ignoring the bad output. Greater awareness of environmental quality has prompted researchers in recent times to rethink their criterion of efficiency measurement. It is now recognized that one must include some penalty for the bad output produced in order to get a measure of the *net* output produced.

An important consideration in this context is: *if some outputs are undesirable, why would not the firm produce the desirable or good output alone without producing the bad output at the same time?* This relates to the concept of *weak disposability.*

One of the critical assumptions about the technology made at the outset was that outputs were freely disposable. Specifically in a 2-output case, it would imply that if the output bundle $y^0=(y_1^0,y_2^0)$ can be produced from some input bundle x^0 then any non-negative output bundle $y^1=(y_1^1,y_2^1)\neq y^0$ can also be produced from x^0 so long as $y_1^1\le y_1^0$ and $y_2^1\le y_2^0$. This would, of course, permit producing the bad output at zero level without reducing the good output! This, however, is not possible because bad outputs are weakly disposable.

As defined by Färe, et al. (2001), the bad output (*b*) and the good (*g*) are weakly disposable if

$$(g^0,b^0)\in T\Rightarrow(kg^0,kb^0)\Rightarrow T\,|\,0\le k\le 1. \qquad (69)$$

That is, the bad output can be reduced only if the good output is reduced proportionately. It is clear that some bad output will necessarily be produced if *any amount of the good output is produced.* Thus, $g=0$ if and only if $b=0$ also. Shephard and Färe (1974) characterized this as *null jointness.* Note that in this interpretation of the relationship between the good and the bad output, the two are produced as joint products. Färe, et al. (2001) assume that while the bad output is weakly disposable (with the good output), the good output, however, is strongly disposable. With weak disposability of the bad output, an empirical estimate of the relevant technology set would be

$$T = \left\{ \begin{aligned} &(g,b;x) : g \le k\sum_j \lambda_j g^j; \\ &b = k\sum_j \lambda_j b^j; x \ge \sum_j \lambda_j x^j; \\ &\sum_j \lambda_j = 1;\ k, \lambda_j \ge 0, (j = 1,2,...,N) \end{aligned} \right\} \quad (70)$$

Of course, under the CRS assumption, the restriction on the sum of the λs does not apply. It can be seen that some of the restrictions in (66) are nonlinear. However, Färe, et al. bypass this nonlinearity problem by setting *k* to unity. Assuming that the criterion of efficiency is simultaneous increase in the good output and decrease in the bad output, relevant graph hyperbolic output oriented DEA problem is[13]

$$\begin{aligned} &\delta^* = \max \delta \\ &s.t. \sum_{j=1}^{N} \lambda_j g_j \ge \delta g_0; \\ &\quad \sum_{j=1}^{N} \lambda_j b_j = \tfrac{1}{\delta} b_0; \\ &\quad \sum_{j=1}^{N} \lambda_j x_{ij} \le x_{i0}, (i = 1,2,...,n); \qquad (71) \\ &\quad \sum_{j=1}^{N} \lambda_j = 1; \\ &\quad \lambda_j \ge 0, j = 1,2,...,N); \delta\, unrestricted. \end{aligned}$$

On the other hand, an output-oriented Directional Distance Function would be

[13] See Färe, Grosskopf, Lovell, and C. Pasurka (1989).

$$
\begin{aligned}
&\vec{D}(x_{10}, x_{20}; g_0, b_0) = \max \beta \\
&s.t. \sum_{j=1}^{N} \lambda_j g_j \geq (1+\beta) g_0; \\
&\quad \sum_{j=1}^{N} \lambda_j b_j = (1-\beta) b_0; \\
&\quad \sum_{j=1}^{N} \lambda_j x_{ij} \leq x_{i0}, (i = 1, 2, ..., n); \\
&\quad \sum_{j=1}^{N} \lambda_j = 1; \\
&\quad \lambda_j \geq 0, j = 1, 2, ..., N); \beta\, unrestricted.
\end{aligned} \tag{72}
$$

A problem with the specification like (71) or (72) is that the technological interdependence between the good and the bad output is not explicitly stated. For example, if the two outputs are joint in the sense that they must always increase or decrease together, how can the good output be treated as strongly disposable while the bad output is only weakly disposable? If a lower amount of power is being generated, where is the smoke coming from? One could argue that lower power generation is the result of inefficient use of resources and the same amount of fossil fuel is being burnt so that the level of pollution is not reduced. In that case, the jointness is directly between fossil fuel and pollution and only indirectly between the good and the bad output. This raises a question of materials balance.

Following Førsund (2009) and Murty and Russell (2010), an alternative interpretation of the bad output would be that it is an unwanted bye product of some input or inputs used for the production of the good output. In agricultural production, for example, fertilizers and chemical pesticides are used along with land, labor, and capital for crop production. But an unwanted consequence of using chemicals is ground water contamination. Thus, the bad output can be reduced only if the polluting inputs are reduced as well. However, to the extent that there is room for input substitution, it may be possible to maintain the crop output level. In this case, weak disposability applies to the bad output and the polluting inputs rather than between the good and the bad output. The underlying production technology involves two separate sub-technologies.

Suppose that there are two outputs *g* (good output) and *b* (bad output) produced from two inputs: x_1 and x_2. Both inputs are used for the production of *g* but *b* is produced from x_2 only. We can think of two production possibility sets: $T_1 = \{(x_1, x_2; g) : g$ can be produced from $(x_1, x_2)\}$ and $T_2 = \{(x_2, b) : b$ can be produced from $x_2\}$.

The usual disposability assumptions are made about the good output and the both inputs in *T₁*. However, the bad output and the offending input (x_2) are assumed to be weakly disposable in T_2.

A an output-oriented Directional Distance Function would be

$$\begin{aligned}
&\vec{D}(x_{10}, x_{20}; g_0, b_0) = \max \beta \\
&s.t. \sum_{j=1}^{N} \lambda_j g_j \geq (1+\beta) g_0; \\
&\sum_{j=1}^{N} \lambda_j b_j = (1-\beta) b_0; \\
&\sum_{j=1}^{N} \lambda_j x_{1j} \leq x_{10}; \qquad\qquad (73) \\
&\sum_{j=1}^{N} \lambda_j x_{2j} = (1-\beta) x_{20}; \\
&\sum_{j=1}^{N} \lambda_j = 1; \\
&\lambda_j \geq 0,\ j = 1, 2, ..., N); \beta\, unrestricted.
\end{aligned}$$

Similarly, a graph hyperbolic (output-oriented) measure of technical efficiency would be $\tau_{Graph} = \frac{1}{\delta^*}$ where

$$\begin{aligned}
&\delta^* = \max \delta \\
&s.t. \sum_{j=1}^{N} \lambda_j g_j \geq \delta g_0; \\
&\sum_{j=1}^{N} \lambda_j b_j = \tfrac{1}{\delta} b_0; \\
&\sum_{j=1}^{N} \lambda_j x_{1j} \leq x_{10}; \qquad\qquad (74) \\
&\sum_{j=1}^{N} \lambda_j x_{2j} = \tfrac{1}{\delta} x_{20}; \\
&\sum_{j=1}^{N} \lambda_j = 1; \\
&\lambda_j \geq 0,\ j = 1, 2, ..., N); \delta\, unrestricted.
\end{aligned}$$

8. DEA with market Prices

There is a widely held belief that DEA should be used only for public sector and non-profit organizations like schools or municipal governments where market prices for outputs are not always available. But for commercial firms which buy and sell their inputs and outputs at observable market prices, one should use parametrically specified econometric models rather than DEA. It should be noted, however, that the cost function relating expenditure to input prices and output quantities, the revenue function relating receipts to output prices and input quantities, or the profit function relating net revenues to prices of inputs and outputs are all derived from the assumptions about the technology and the objectives of the firm. Econometrics and DEA are two alternative methods of calibrating the relationship between the prices, quantities, expenses, and revenues as relevant in a particular problem. Choice between the two alternative techniques should not depend on the availability or lack of information about market prices of inputs and outputs.

DEA for Cost Minimization

Consider a producer using the input bundle x^0 to produce the output bundle y^0. Further assume that the market price vector of the inputs is w^0 and the firm is a price taker in the input market. Then its actual cost is $C_0 = w^{0\prime}x^0$. Clearly, because y^0 is being produced from $x^0, (x^0, y^0)$ is a feasible input-output combination. That is $(x^0, y^0) \in \mathrm{T}$. the question is whether $C_0 = w^{0\prime}x^0$ is the *minimum* cost of producing the output bundle y^0. The cost minimization problem of the firm is to

$$\min w^0{}'x : (x, y^0) \in T. \qquad (75)$$

Suppose x^* is the cost minimizing input bundle and $C^* = w^{0\prime}x^*$ is the minimum cost. Given a reference technology, this minimum cost will depend on both the input price vector w^0 and the target output bundle y^0 and can be expressed as

$$C(w^0, y^0) = \min w^{0\prime}x : (x, y^0) \in T. \qquad (76)$$

In production economics, the minimum cost function $C^* = C(w, y)$ is known as the dual cost function.

Using the free disposal convex hull, *S* (from (6) above) as an estimate of the production possibility set, the DEA problem for cost minimization can be set up as

$$\begin{aligned}
&C^* = \min w^{0\prime} x \\
&s.t. \sum_{j=1}^{N} \lambda_j x^j \leq x; \\
&\sum_{j=1}^{N} \lambda_j y^j \geq y^0; \qquad\qquad\qquad (77) \\
&\sum_{j=1}^{N} \lambda_j = 1; \\
&x \geq 0; \lambda_j \geq 0 \,(j = 1, 2, ..., N).
\end{aligned}$$

If CRS is assumed, the constrained $\sum_{j=1}^{N} \lambda_j = 1$ is excluded. Like the λ_js the optimal input vector x also is a (vector of) choice variable(s) in this LP problem. Note that at the optimal solution there cannot be any slacks in any of the input constraints. To see that, define

$$\sum_{j=1}^{N} \lambda_j x^j = \bar{x}; \sum_{j=1}^{N} \lambda_j y^j = \bar{y}.$$

.

By convexity, $(\bar{x}, \bar{y}) \in T.$ Further, by free disposability of outputs, $\bar{y} \geq y^0 \Rightarrow (\bar{x}, y^0) \in T.$ Hence, the minimum cost cannot be any higher than $w^{0\prime}\bar{x}.$ But if there is any input slack at the optimal solution (x^*), $w^{0\prime}x^0 > w^{0\prime}\bar{x}.$ In that case, $w^{0\prime}x^*$ cannot be the optimal solution of the cost minimization problem in (73).

The (overall) cost efficiency of the firm can be measured as

$$\gamma = \frac{C(w^0, y^0)}{C_0} = \frac{w^{0\prime}x^*}{w^{0\prime}x^0}. \qquad (78)$$

Farrell (1957) provides an interesting decomposition of the overall efficiency into two distinct components denoting technical and allocative efficiency.

Consider the observed input-output bundle of the firm, (x^0, y^0) and its input-oriented technical efficiency: $\tau_x(x^0, y^0) = \beta = \min \theta : (\theta x^0, y^0) \in T.$ Because $(x^0, y^0) \in \mathrm{T}, \beta \leq 1.$ Now define $x_t^0 = \beta x^0.$ It is the technically efficient projection of the observed input bundle $x^0.$ The cost of this technically efficient input bundle is

$$C_t = w^{0\prime}x_t^0 = \beta w^{0\prime}x^0 = \beta C_0.$$

Obviously, $\frac{C_t}{C_0} = \beta$ is the technical efficiency of the firm. Next compare $C_t = w^{0\prime} x_t^0$ with the minimum cost $C^* = w^{0\prime} x^*$. It follows from the definition of a minimum, that $C^* \leq w^{0\prime} x$ over all input bundles x so long as $(x, y^0) \in T$. Because the bundle x_t^0 is one such bundle, $C^* = w^{0\prime} x^* \leq w^{0\prime} x_t^0 = C_t$. Farrell defined the ratio

$$\alpha \equiv \frac{w^{0\prime} x^*}{w^{0\prime} x_t^0} \text{ as allocative efficiency.}$$

Hence we have the decomposition

$$\frac{C^*}{C_0} = \left(\frac{C_t}{C_0} \right) . \left(\frac{C^*}{C_t} \right) \qquad (79)$$

Or,

$$\gamma = (\beta).(\alpha). \qquad (80)$$

As explained above, each of the three ratios, α, β, and γ lies between 0 and 1. The measure of technical efficiency (*β*) shows the potential reduction in cost by proportional reduction in all inputs without changing the output. By contrast, allocative efficiency (*α*) measures the reduction in cost by changing the input mix and substituting a relatively less expensive input for another which is (relatively) more expensive. The overall cost efficiency (*γ*) reflects the potential for cost reduction by scaling down the input bundle to the extent possible and then selecting a different input mix $\left(\frac{C_t}{C_0} \right)$ to take advantage of input substitution.

Profit Maximization

Finally one can consider the problem of a profit-maximizing firm in a competitive market producing *m* outputs. The output price vector $p >> 0$ is determined by market demand and supply and is not within the control of the firm. The firm merely selects the optimal input-output bundle that is feasible and maximizes the difference between revenue and cost at the applicable market prices of outputs and inputs. Thus, conceptually, the maximum profit is

$$\pi^* = \max p'y - w'x : (x, y) \in T. \qquad (81)$$

With reference to the empirically constructed set *S*, the relevant DEA problem becomes

$$\begin{aligned} &\max \pi = p'y - w'x \\ &s.t. \ \sum_{j=1}^{N} \lambda_j y^j \geq y; \\ &\quad \sum_{j=1}^{N} \lambda_j x^j \geq x; \qquad\qquad (82) \\ &\quad \sum_{j=1}^{N} \lambda_j = 1; \lambda_j \geq 0 \, (j = 1, 2, ..., N). \end{aligned}$$

Note that in this problem, the right hand sides of both the input and output constraints are themselves choice variables. Another important point is that for the profit maximization problem (without any other constraint), *one must allow variable returns to scale*. If CRS is assumed, for every feasible (x^0, y^0) that yields the profit $\pi^0 = p'y^0 - w'x^0$, (t x^0, t y^0) is also feasible. Hence, the profit will either unbounded or 0. In fact, locally diminishing returns must hold at the optimal point (x^*, y^*) on the frontier. Otherwise, if (kx^*, ky^*) is feasible for some k >1, it is possible to increase the profit to

$$\pi = p'(ky^*) - w'(kx^*) = k(p'y^* - w'x^*)$$

and (x^*, y^*) cannot be the profit-maximizing bundle.

A common measure of profit efficiency is

$$\rho = \frac{\pi^0}{\pi^*} \qquad (83)$$

where (x^0, y^0) is the actual input-output of the firm under evaluation and $\pi^0 = p'y^0 - w'x^0$ is its actual profit. It should be remembered, however, that if actual profit is negative, the efficiency falls below zero. Additionally, if the maximum profit is also negative, the ratio exceeds unity. This may appear strange.

However, from the stand point of pure economic theory, because *all* inputs and outputs are being freely chosen, we are considering what is known as the long run profit maximization problem. If no input-output bundle yields a positive profit, the firm should have the option to shut down and earn zero profit. In the LP problem a zero input-output bundle can be selected only if all λs are set to 0. But that would violate the summation constraint on the λs. So, if a firm is earning negative profit but is still in business, that is not a long run solution in the strict sense and there must be some constraints that keep it from closing down.

9. Summing up

Although it was introduced as an optimization problem in the OR/MS literature, DEA has grown into a nonparametric alternative to stochastic frontier analysis (SFA) for measurement of production efficiency. This paper provides the neoclassical production economics behind the different formulations of DEA for measurement of technical an, scale, cost, and profit efficiency. Given the limited scope of the overview many important issues could not be addressed. Most important among them are (i) measurement and decomposition of total factor productivity growth over time, (ii) the role of exogenous (or contextual) factors in efficiency measurement, and (iii) bootstrapping for generating confidence intervals of DEA efficiency measures. The more ambitious reader should refer to Ray (2004) and Cooper, Seiford, and Tone (2000).

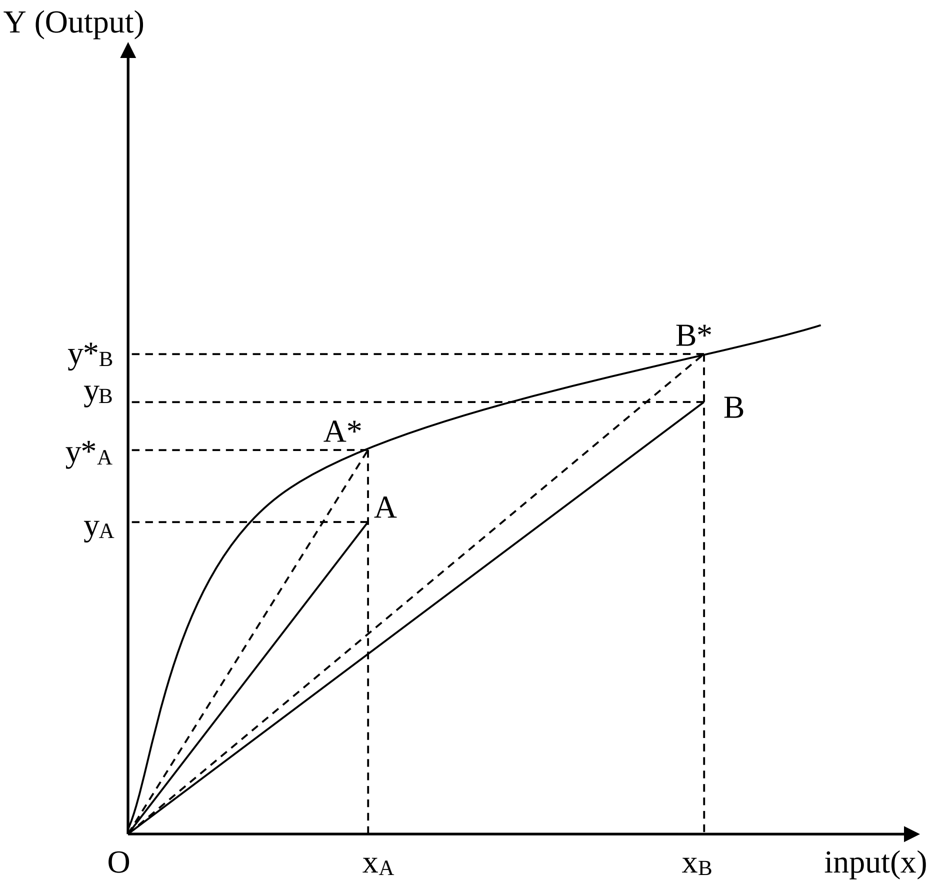

Figure 1a Output oriented Technical Efficiency

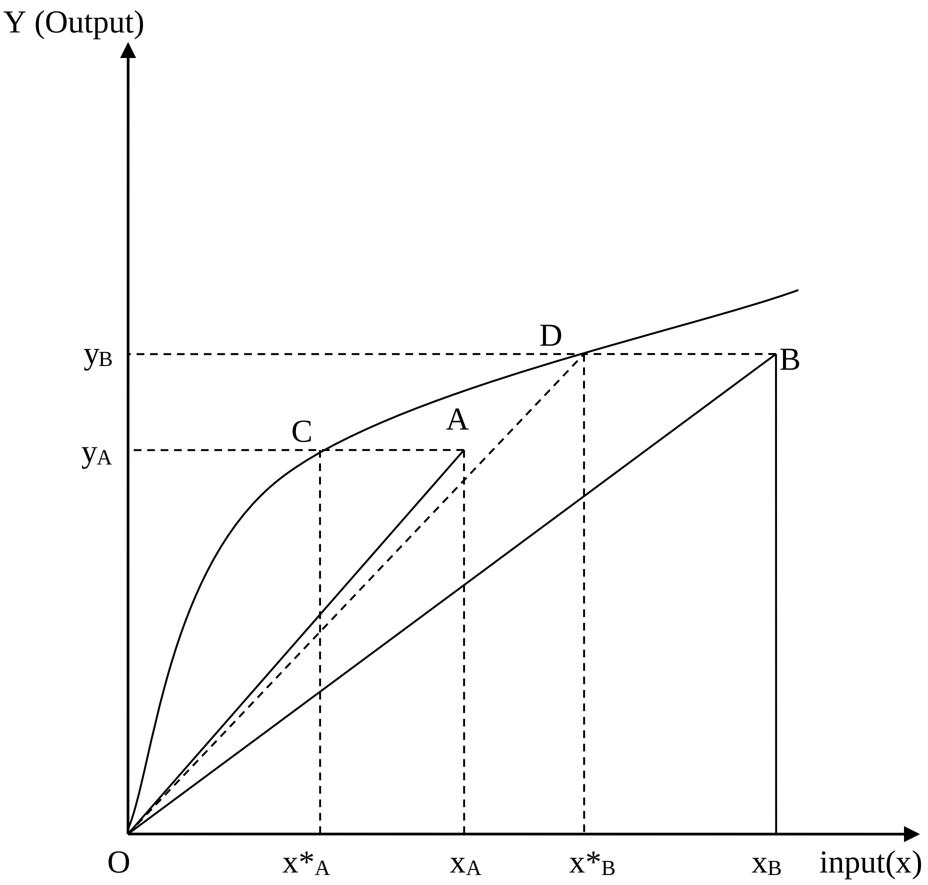

Figure 1b Input oriented Technical Efficiency

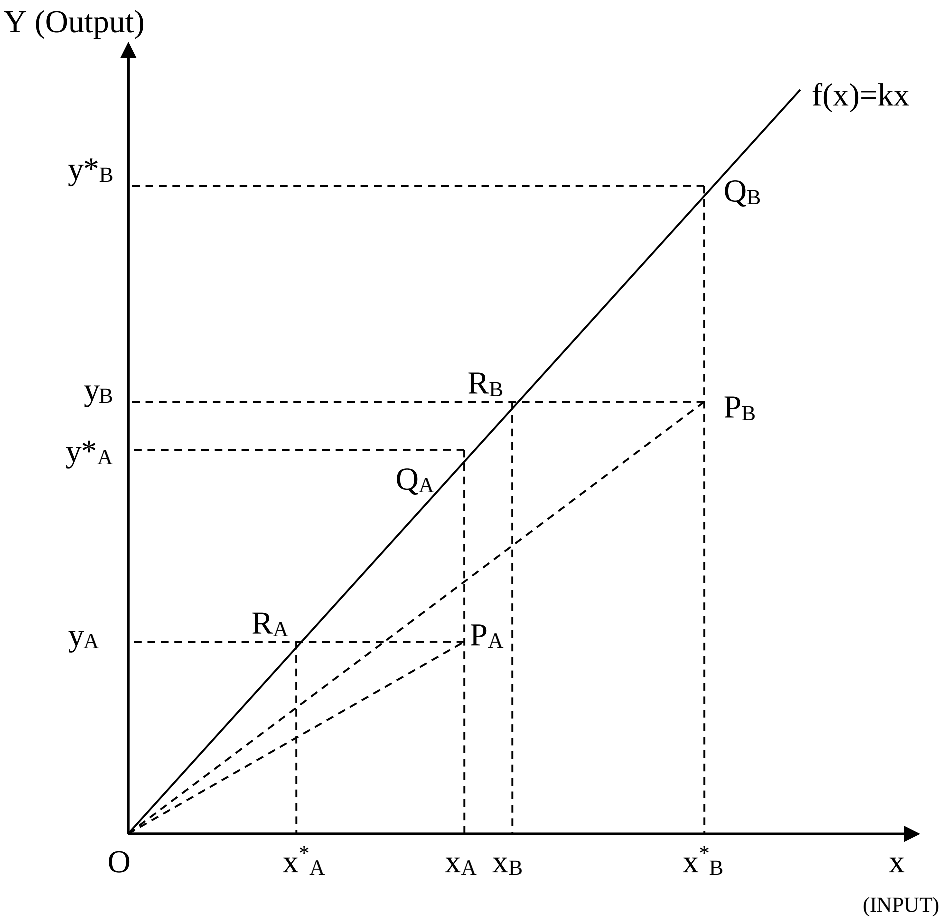

Figure 1c Technical Efficiency under Constant Returns to Scale

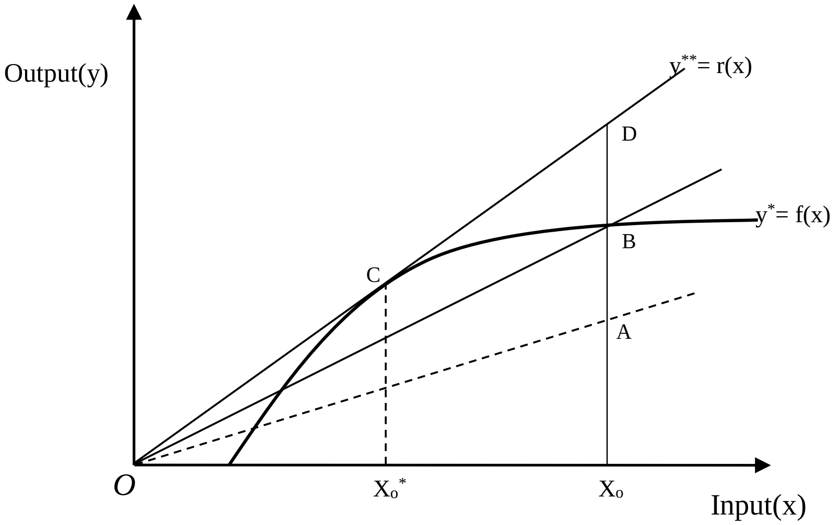

Figure 2 Scale Efficiency

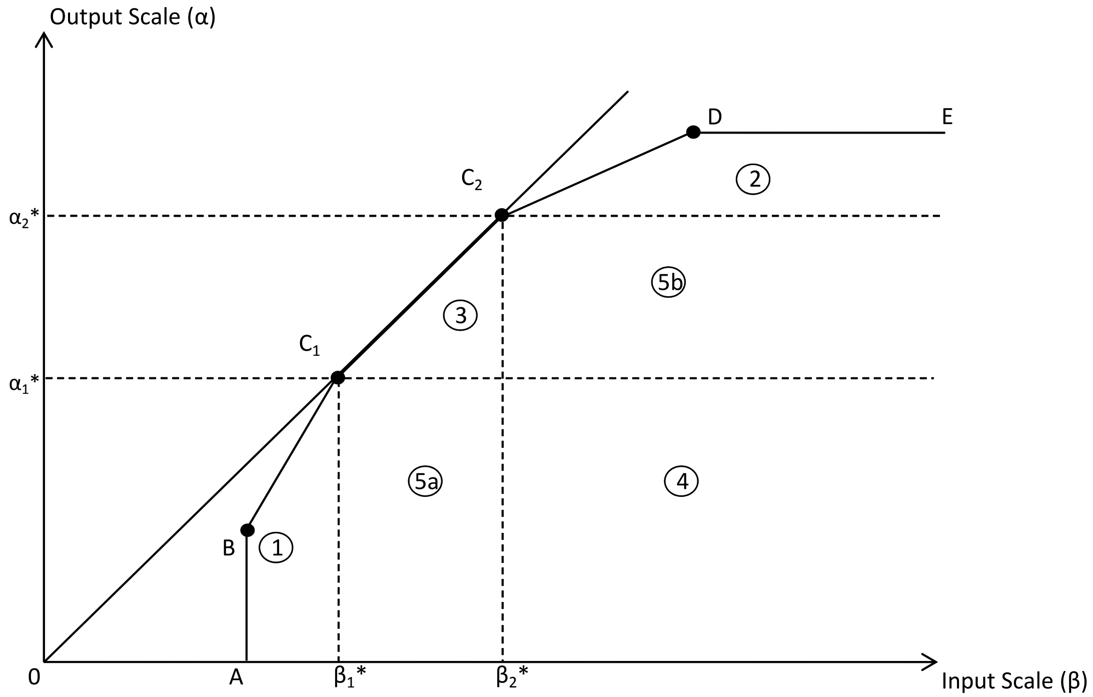

Figure 3: Multiple MPSS & Regions of Increasing, Constant, Decreasing, and Ambiguous Returns to Scale

References:

Aigner, D.J., C.A.K. Lovell, and P. Schmidt (1977) " Formulation and Estimation of Stochastic Frontier Production Function Models"; *Journal of Econometrics,* 6:1, 21-37.

Ali, A.I. and L.M. Seiford (1990) "Translation Invariance in Data Envelopment Analysis"; *Operations Research Letters,* 9, 403-405.

Banker, R.D. (1984), "Estimating the Most Productive Scale Size Using Data Envelopment Analysis", *European Journal of Operational Research* 17: 1 (July) 35-44.

Banker, R.D., H. Chang., and W.W. Cooper (1996) "Equivalence and Implementation of Alternative Methods of Determining Returns to Scale in Data Envelopment Analysis", *European Journal of Operational Research,* 89, 583-585.

Banker, R.D., A. Charnes, and W.W. Cooper (1984), "Some Models for Estimating Technical and Scale Inefficiencies in Data Envelopment Analysis," *Management Science,* 30:9 (September), 1078-92.

Banker, R.D., W.W. Cooper, L. M. Seiford, R.M. Thrall, and J. Zhu (2004) "Returns to Scale in different DEA Models"; *European Journal of Operational Research* 154: 345-362.

Banker, R.D. and R.M. Thrall (1992) " Estimating Most Productive Scale Size using Data Envelopment Analysis", *European Journal of Operational Research,62, 74-84.*

Byrnes, P., R. Färe, and S. Grosskopf (1984), "Measuring Productive Efficiency: An Application to Illinois Strip Mines", *Management Science* 30:6 (June) 671-681.

Chambers, R.G., Y. Chung, and R. Färe (1996) Benefit and Distance Functions", *Journal of Economic Theory,* 70 (August 1996), 407-19.

Charnes, A. and W.W. Cooper (1968) "Programming with Linear Fractional Functionals", *Naval Research Logistics Quarterly* 15; 517-522.

Charnes, A., W.W. Cooper, and E. Rhodes (1978) "Measuring the Efficiency of Decision Making Units," *European Journal of Operational Research* 2:6 (November), 429-44.

Charnes, A., W.W. Cooper, and E. Rhodes (1979) " Short Communication: Measuring the Efficiency of Decision Making Units," *European Journal of Operational Research* 3:4, 339

Cooper, W.W., L. Seiford and K. Tone.(2000) *Data Envelopment Analysis: A Comprehensive Text with Uses, Example Applications, References and DEA-Solver Software*. (Norwell, Mass: Kluwer Academic Publishers.)

Cooper, W.W., R.G. Thompson, and R.M. Thrall (1996) "Introduction: Extensions and New Developments in DEA"; *Annals of Operations Research*, 66, pp 3-45.

Debreu, G. (1951), "The Coefficient of Resource Utilization," *Econometrica* 19:3 (July), 273-92.

Färe, R. and S. Grosskopf (2000) "Theory and Application of Directional Distance Functions"; *Journal of Productivity Analysis* 13:93-103..

Färe, R. and C.A.K. Lovell (1978), "Measuring the Technical Efficiency of Production," *Journal of Economic Theory* 19:1 (October), 150-62.

Färe, R., S. Grosskopf, and C.A.K. Lovell (1985) *The Measurement of Efficiency of Production* Boston:

Kluwer-Nijhoff.

Färe, R., S. Grosskopf, and C.A.K. Lovell (1994) *Production Frontiers* Cambridge: Cambridge University Press.

Färe, R., S. Grosskopf, and C.A.K. Lovell, and C. Pasurka (1989) "Multilateral Productivity Comparisons when Some Outputs are Undesirable: A Non-parametric Approach", *Review of Economics and Statistics,*71:1 (February) 90-98.

Farrell, M.J. (1957), "The Measurement of Technical Efficiency,", *Journal of the Royal Statistical Society* Series A, General, 120, Part 3, 253-81.

Førsund, F. and L.Hjalmarsson (1979) "Generalized Farrell Measures of Efficiency: An Application to Milk Processing in Swedish Dairy Plants"; *The Economic Journal;*(89) 354, 294-315.

Førsund, F. "Good Modelling of Bad Outputs: Pollution and Multiple-Output Production"; *International Review of Environmental and Resource Economics*; Vol 3, Issue 1, 009, pp 1-38)

Førsund, F. (2013) "Weight restrictions in DEA: misplaced emphasis?"; *Journal of Productivity Analysis* 40:271–283.

Frisch, R. (1965) *Theory of Production.* Chicago: Rand McNally and Company.

Kerstens, K. and P. Vanden Eeckaut (1999) "Estimating Returns to Scale Using Nonparametric Deterministic Technologies: A New Method Based on Goodness-of-Fit"; *European Journal of Operational Research,* 113(1), 206-214.

Kumbhakar, S. and C.A.K. Lovell (2000) *Stochastic Frontier Analysis* (New York: Cambridge University Press).

Lovell, C.A.K. and J.T. Pastor (1995) "Units Invariant and Translation Invariant DEA Models"; *Operations Research Letters,* 18, 147-151.

Luenberger, D.G. (1992) "Benefit Functions and Duality", *Journal of Mathematical Economics,* 21, 115-145.

Meeusen, W. and J. van den Broeck (1977) "Efficiency Estimation from Cobb-Douglas Production Functions with Composed Errors"; *International Economic Review,* 18:2; 435-444.

Murty, S., R. Russell, and S. Levkoff. (2012) "On modeling pollution-generating technologies." *Journal of Environmental Economics and Management* 64 (2012) 117–135

Pastor, J.T., J.L. Ruiz, and I. Sirvent (1999) "An Enhanced DEA Russell-Graph Efficiency Measure", *European Journal of Operational Research,* 115, 596-607.

Podinovski, V. (2004) "Local and Global Returns to Scale in performance Measurement"; *Journal of the Operations Research Society* (2004) 55, 170-178.

Portela, M.C.A.S., and Thanassoulis, E. (2005) Profitability of a sample of Portuguese bank branches and its decomposition into technical and allocative components, *European Journal of Operational Research* 162/3, 850–866.

Ray, S.C. (2004) *Data Envelopment Analysis: Theory and Techniques for Economics and Operations Research* (New York: Cambridge University Press).

Ray, S.C. (2010) "A One-Step Procedure for Returns to Scale Classification of Decision Making Units in Data Envelopment Analysis" *University of Connecticut Economics Working Paper* 2010-07.

Ray, S.C. and Y. Jeon (2009) "Reputation and efficiency: A non-parametric assessment of America's top-rated MBA programs"; *European Journal of Operational Research* 189 (2008) 245–268

Ray, S.C. and A. Ghose (2014) "Production efficiency in Indian agriculture: An assessment of the post green revolution years"; *Omega* 44(2014)58–69

Seiford, L. and J. Zhu. "An investigation of returns to scale in data envelopment analysis"; *Omega, Int. J. Mgmt. Sci.* 27 (1999) 1-11

Shephard, R. W. (1953), *Cost and Production Functions* Princeton: Princeton University Press.

Shephard, R.W., and R. Färe (1974), The Law of Diminishing Returns, *Zeitschrift für Nationalökonomie* 34, 69–90.

Tone, K. (2001) A slacks-based measure of efficiency in data envelopment analysis, *European Journal of Operational Research* 130, 498–509

Zhu, J.(2003) *Quantitative Models for Performance Evaluation and Benchmarking: Data Envelopment Analysis with Spreadsheets and DEA Excel Solver* (Boston: Kluwer Academic Press).