

Journal of Econometrics 30 (1985) 91-107. North-Holland

FOUNDATIONS OF DATA ENVELOPMENT ANALYSIS FOR PARETO-KOOPMANS EFFICIENT EMPIRICAL PRODUCTION FUNCTIONS*

A. CHARNES, W.W. COOPER, B. GOLANY and L. SEIFORD

The University of Texas, Austin, TX 78712, USA

J. STUTZ

University of Miami, Coral Gables, FL 00000, USA

The construction and analysis of Pareto-efficient frontier production functions by a new Data Envelopment Analysis method is presented in the context of new theoretical characterizations of the inherent structure and capabilities of such empirical production functions. Contrasts and connections with other developments, including solutions of some remaining problems, are made re aspects such as informatics, economies of scale, isotonicity and non-concavity, discretionary and non-discretionary inputs, piecewise linearity, partial derivatives and Cobb-Douglas properties of the functions. Non-Archimedean constructs are *not* required.

1. Introduction

Classically, the economic theory of production is heavily based on the conceptual use of the Pareto-efficiency (or Pareto-optimal) frontier of production possibility sets to define *the* production function. The work of Shephard (1953, 1970), under restrictions on the mathematical structure of production possibility sets and cost relations, developed an elegant 'transform' theory between production aspects and cost aspects as in Charnes, Cooper and Schinnar (1982). This was applied to various classes of explicitly given parametric functional forms, and problems of statistical estimation of parameters from data were considered in classical statistical contexts especially by successors like Afriat (1972), Aigner, Lovell and Schmidt (1977), and Försund and Hjalmarsson (1979).¹ These efforts were almost exclusively for single-output functions.

Farrell (1957), partly responding to the inadequacies of separate indices of labor productivity, capital productivity, etc., undertook what he referred to as

*This research was partly supported by NSF Grant SES-8408134 and USARI Contract MDA903-83-K-0312 with the Center for Cybernetic Studies, The University of Texas at Austin. Reproduction in whole or in part is permitted for any purpose of the United States Government.

¹The need for some method of estimating frontier functions had been pointed out as early as 1935 by R. Frisch (1935) in his statistical study of chocolate production in France.

an 'activity analysis' approach² that could deal more adequately with the problem. Possibly because of the limitations of the elaborate matrix inversion routines he was employing, Farrell confined his numerical examples and discussion to single-output situations, although he did formulate the multiple-output case. If anything, the need for being able to deal directly with multiple outputs has grown ever more urgent with the continuing growth of the not-for-profit and service sectors in the United States and other 'western type' economies. As we have elsewhere observed,³ one can otherwise do almost nothing with important parts of not-for-profit and governmental activities without becoming involved in arbitrary weighting schemes. Even attempts at measuring and evaluating 'total factor productivity' will fall short of what is required since this is inherently a single-output concept. Farrell was right, therefore, in re-orienting the direction of development toward the various types of efficiencies that he referred to as 'technical', 'scale', and 'allocative efficiencies'.

Building on the individual firm or country (= Decision Making Unit)⁴ evaluations of Farrell and the engineering ratio idea of efficiency measure for a single-input, single-output efficiency analysis in its managerial aspects and its constructible extensions to multi-input, multi-output situations was initiated by Charnes, Cooper and Rhodes (1978, 1981). Subsequent extensions and elaborations by the former pair with other students and colleagues [Charnes, Cooper, Lewin, Morey and Rousseau (forthcoming), Charnes, Cooper, Seiford and Stutz (1982, 1983), and Charnes, Cooper and Sherman (forthcoming)] were made with more detailed attention to classical economic aspects and deeper analysis of the production function side of the mathematical duality structure and Data Envelopment Analysis first presented in the original CCR work. The CCR ratio measures and the variants of Farrell, Shephard, Färe et al. on the dual linear programming side require, however, non-Archimedean constructs for rigorous theory and usage. Their solution methods also do not easily provide important needed properties of their associated empirical production functions.

Thus, in this paper we introduce as basic the idea of Pareto-optimality with respect to an empirically defined production possibility set. We characterize the mathematical structures permitted under our minimal assumptions and contrast these with work by others. Properties such as isotonicity, non-concavity, economies of scale, piece-wise linearity, Cobb-Douglas forms, discretionary and non-discretionary inputs are treated through a new Data Envelopment

²In the sense of the original work by T.C. Koopmans that is described in Chapter IX of Charnes and Cooper (1980) where it was also accorded an explicit linear programming formulation and interpretation.

³See Banker, Bowlin, Charnes and Cooper (in process).

⁴This term was introduced in Charnes, Cooper and Rhodes (1978) and contracted to DMU because it was recognized that something like it was needed to refer to public sector organizations.

deal more adequately with the the elaborate matrix inversion his numerical examples and he did formulate the multiple- to deal directly with multiple he continuing growth of the states and other 'western type' one can otherwise do almost and governmental activities ing schemes. Even attempts at 'ity' will fall short of what is t concept. Farrell was right, nent toward the various types 'cale', and 'allocative efficien-

(= Decision Making Unit)⁴ ea of efficiency measure for a s managerial aspects and its tput situations was initiated Subsequent extensions and its a colleagues [Charnes, g), Charnes, Cooper, Seiford 'herman (forthcoming)] were onomic aspects and deeper thematical duality structure the original CCR work. The hephard, Färe et al. on the on-Archimedean constructs ethods also do not easily iated empirical production

a of Pareto-optimality with ibility set. We characterize minimal assumptions and as isotonicity, non-concav- Douglas forms, discretion- a new Data Envelopment

cribed in Chapter IX of Charnes r programming formulation and

1978) and contracted to DMU r to public sector organizations.

Analysis method and informatics which permit a constructive development of an empirical production function and its partial derivatives without loss of efficiency analysis or use of non-Archimedean field extensions.

2. Empirical function setting and generation

By an 'empirical' function we shall mean a vector function whose values are known at a finite number of points and whose values at other points in its domain are given by linear (usually convex) combinations of values at known points. The points in the domain are 'inputs', the component values of the vector function 'output'. We shall assume that inputs are so chosen that convex combinations of input values for each input are meaningful input values. We assume this for output values as well.

In efficiency analysis, observations are generated by a finite number of 'DMUs', or 'productive', or 'response' units, all of which have the same inputs and outputs. A relative efficiency rating is to be obtained for each unit. Typically, observations over time will be made of each unit and the results of efficiency analyses will be employed to assist in managing each of the units. We assume n units, s outputs and m inputs. The observed values are to be non-negative (sometimes positive) numbers.

We shall employ the notation X_j, Y_j for respectively the observed vectors of the inputs and outputs of the j th DMU. By X and Y we shall mean the matrices whose n column vectors are respectively those of the n DMUs. We use x, y, λ and s to denote column vectors, also with e the column vector of ones. We use 'T' as a superscript to denote the transpose, e.g. e^T is the row vector of ones. Point sets are denoted by (capital) script letters. The inequality symbol between vectors means that inequality holds for each component.

3. A hypograph empirical production possibility set

Given the (empirical) points (X_j, Y_j) , $j = 1, \dots, n$, with $(m \times 1)$ 'input' vectors $X_j \geq 0$ and $(s \times 1)$ 'output' vectors $Y_j \geq 0$, we define the 'empirical production set' $\mathcal{P}_E$ to be the convex hull of these points, i.e.,

$$\mathcal{P}_E \triangleq \left\{ (x, y) : x = \sum_{j=1}^n X_j \mu_j, y = \sum_{j=1}^n Y_j \mu_j, \forall \mu_j \geq 0, \sum_j \mu_j = 1 \right\}. \quad (3.1)$$

We extend it to our 'empirical production possibility set' $\mathcal{Q}_E$ by adding to $\mathcal{P}_E$ all points with inputs in $\mathcal{P}_E$ and outputs not greater than some output in $\mathcal{P}_E$, i.e.,

$$\mathcal{Q}_E \triangleq \{ (x, y) : x = \bar{x}, y \leq \bar{y} \text{ for some } (\bar{x}, \bar{y}) \in \mathcal{P}_E \}. \quad (3.2)$$

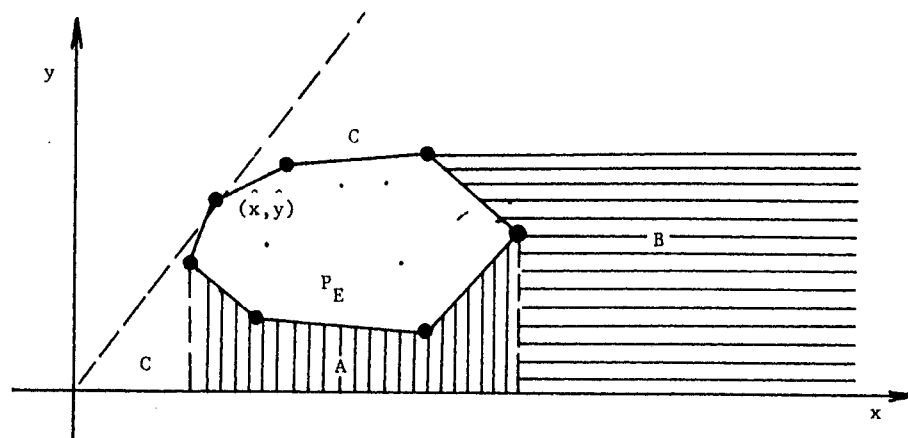


Fig. 1

Note that $\mathcal{Q}_E$ is contained in (e.g., is smaller than) *every* production possibility set heretofore employed, i.e., those studies by Farrell (1957), Shephard (1970), Banker, Charnes and Cooper (1984), Fare and Lovell (1978), etc. We also use fewer axioms than the others, including even Banker (1984) who, up to this point in *journal* publication, had used the most parsimonious axiom system to characterize the production possibility set of DEA.⁵ The Farrell, Shephard, Färe sets are (truncated) cones; the BCC set (when not also a cone) adds to $\mathcal{Q}_E$ the set

$$\{(x, y): x \geq \bar{x}, y = \bar{y} \text{ for some } (\bar{x}, \bar{y}) \in \mathcal{Q}_E\}.$$

These relations may be visualized in the schematic plot of fig. 1, where $\mathcal{Q}_E = \mathcal{P}_E \cup \mathcal{A}$, the BCC set is $\mathcal{Q}_E \cup \mathcal{B}$, and the Farrell, Shephard, Färe set is $\mathcal{Q}_E \cup \mathcal{B} \cup \mathcal{C}$.

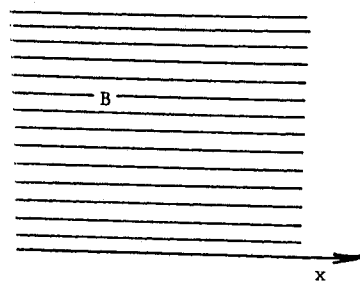
Let $\mathcal{P}_E^*, \mathcal{Q}_E^*$ denote the sets corresponding to $\mathcal{P}_E$ and $\mathcal{Q}_E$ when only the output y_i is the ordinate. Evidently a frontier function $f_i(x)$ is determined by

$$f_i(x) = \max y_i \text{ for } (x, y_i) \in \mathcal{Q}_E^*. \quad (3.3)$$

Then:

Theorem 1. $\mathcal{Q}_E^*$ is the hypograph of $f_i(x)$ over $\{x: (x, y) \in \mathcal{Q}_E\}$.

⁵Our usage antedates his since we used this in the Ph.D. and other work with Rhodes (the 'R' of CCR).



every production possibility
rell (1957), Shephard (1970),
vell (1978), etc. We also use
nker (1984) who, up to this
rsin... our axiom system to
A.⁵ The Farrell, Shephard,
not also a cone) adds to $\mathcal{Q}_E$

}.
(3.3)

natic plot of fig. 1, where
rell, Shephard, Färe set is

$\mathcal{Q}_E$ and $\mathcal{Q}_E$ when only the
on $f_i(x)$ is determined by

(3.3)

$(x, y) \in \mathcal{Q}_E$.

er work with Rhodes (the 'R' of

Proof. The hypograph H_i of $f_i(x)$ is the set

$$H_i \triangleq \{(x, y) : y_i \leq f_i(x), (x, y) \in \mathcal{Q}_E\}. \quad \blacksquare$$

Let $\mathcal{D}_E$ denote $\{x : (x, y) \in \mathcal{Q}_E\}$. It is the domain (the input set) of our empirical frontier functions.

Theorem 2. $f_i(x)$ is a concave, piecewise linear function on $\mathcal{D}_E$.

Proof. A necessary and sufficient condition for $f_i(x)$ to be concave is that its hypograph is a convex set [cf. Rockafellar (1970) or Fenchel (1953)]. The piecewise linearity also follows from the construction of $\mathcal{Q}_E$ by all convex combinations of the empirical points (X_j, Y_j) , $j = 1, \dots, n$. $\blacksquare$

We further observe explicitly that *no use* whatever has been made of *non-negativity* of input and output values in the sets, functions or proof of Theorems 1 and 2. Therefore, they hold without this restriction – a fact we shall employ elsewhere.

Also, *no* assumptions have been made about the properties of *any* underlying function, or function hypograph, from which the (X_j, Y_j) of our empirical construct may be considered samples. Theorem 2 shows, therefore, that any *empirical* (maximum) frontier function is the 'concave cap' function of its graph.

4. The empirical Pareto-efficient production function

A Pareto-efficient (minimum) point for a finite set of functions $g_1(x), \dots, g_K(x)$ is a point x^* such that there is no other point x in the domain of these functions such that

$$g_k(x) \leq g_k(x^*), \quad k = 1, \dots, K, \quad (4.1)$$

with at least one strict inequality. Charnes and Cooper (1962, ch. IX) showed that x^* is Pareto-efficient iff x^* is an optimal solution to the mathematical (goal) program⁶

$$\min_x \sum_{k=1}^K g_k(x) \quad \text{subject to} \quad g_k(x) \leq g_k(x^*), \quad k = 1, \dots, K. \quad (4.2)$$

This was employed by Ben-Israel, Ben-Tal and Charnes (1977) to develop the

⁶For a formal definition of goal programming and some of its history, see Charnes and Cooper (1977).

currently strongest necessary and sufficient conditions for a Pareto-minimum in convex programming.

Utilizing (4.2) we can now define and construct, im(or ex-)plicitly, the Pareto-efficient empirical (frontier) production function. Because of Koopmans' work in this area [see Charnes and Cooper (1961)], we shall use interchangeably the designations 'Pareto-efficient' and 'Pareto-Koopmans efficient', etc. in this paper. Other usages of (4.2) to generalizations such as the 'functional efficiency' of Charnes and Cooper (1961) will not be developed here.

First, by (4.2), the Pareto-efficient points among our n empirical points can be determined. The empirical Pareto-efficient function is then defined on the convex hull of their inputs by convex combinations of the 'output' values. Note that the convex hull of the Pareto-efficient points might not include all of $\mathcal{P}_E$ since only the doubled line portion of the frontier corresponds to Pareto-efficient points.

Since for efficient production we wish to maximize on outputs while minimizing on inputs, our relevant $g_k(x)$ include both outputs and inputs, e.g.,

$$\begin{aligned} -g_k(x) &\triangleq y_k, & 1 \leq k \leq s, \\ &\triangleq -x_i, & k = s + i, \quad i = 1, \dots, m, \quad \text{for } (x, y) \in \mathcal{Q}_E, \end{aligned} \quad (4.3)$$

For the optimization in (4.2) we clearly need only consider $(x, y) \in \mathcal{P}_E$ rather than $\mathcal{Q}_E$. Thus the constraint inequalities in (4.2) are for a test point (x^*, y^*) :

$$y \geq y^*, \quad x \leq x^*, \quad (4.4)$$

and we have, since these are the envelopment constraints of Data Envelopment Analysis for an observed input vector x^* and corresponding output vector y^* :

Theorem 3. The envelopment constraints of Data Envelopment Analysis in production analysis are the Charnes-Cooper constraints of (4.2) for testing Pareto-Koopmans efficiency of an empirical production point.

In no way is what we call 'Data Envelopment Analysis' restricted to linear constant returns to scale functions or to truncated cone domains. Evidently via (4.2), Data Envelopment Analysis applies to much more general convex functions, function domains and other situations than the current *empirical* production function one.

To test an empirical 'input-output' point (X_0, Y_0) for Pareto-Koopmans efficiency, the Charnes-Cooper test of (4.2) becomes

$$\min_{\lambda, s^+, s^-} -e^T Y \lambda + e^T X \lambda,$$

empirical production functions

conditions for a Pareto-minimum

construct, im(or ex-)plicitly, function. Because of d Cooper (1961)], we shall use it' and 'Pareto-Koopmans efficient' generalizations such as the (1961) will not be developed

ng our n empirical points can nction is then defined on the ns of the 'output' values. Note s might not include all of $\mathcal{P}_E$ er corresponds to Pareto-effi-

imize on outputs while mini- n outputs and inputs, e.g.,

$$t, \text{ for } (x, y) \in \mathcal{Q}_E, \quad (4.3)$$

consider $(x, y) \in \mathcal{P}_E$ rather re for a test point (x^*, y^*) :

$$(4.4)$$

aint: Data Envelopment sponding output vector y^* :

Envelopment Analysis in raints of (4.2) for testing n point.

alysis' restricted to linear ne domains. Evidently via ch more general convex an the current empirical

) for Pareto-Koopmans

subject to

$$Y\lambda - s^+ = Y_0, \quad -X\lambda - s^- = -X_0, \quad e^T\lambda = 1, \quad (4.5)$$

$$\lambda, s^+, s^- \geq 0,$$

where

$$X \triangleq [X_1, \dots, X_n], \quad Y \triangleq [Y_1, \dots, Y_n].$$

Since $-e^T(Y\lambda - Y_0) + e^T(X\lambda - X_0)$ is an equivalent functional (it differs from the above one only by a constant), we can rewrite the problem for convenience in later comparisons as

$$\min_{\lambda, s^+, s^-} -e^Ts^+ - e^Ts^-,$$

subject to

$$Y\lambda - s^+ = Y_0, \quad -X\lambda - s^- = -X_0, \quad e^T\lambda = 1, \quad (4.6)$$

$$\lambda, s^+, s^- \geq 0.$$

Here, $\min -e^Ts^+ - e^Ts^- = -e^Ts^+ - e^Ts^- = 0$ if and only if Y_0, X_0 is Pareto-Koopmans efficient.⁷ This is the new DEA form for obtaining and characterizing the production possibility set $\mathcal{Q}_E$ via $\mathcal{P}_E$. The linear program (4.6) maximizes the l_1 -distance of a point in $\mathcal{P}_E$ to (X_0, Y_0) . We solve (4.6) for all n DMUs considered as (X_0, Y_0) . From these we get the efficient and non-efficient points and can construct as desired the points of $\mathcal{P}_E$ by convex combinations. As we shall see later, other variations of $\mathcal{Q}_E$ can be accommodated easily by simple modifications of or additions to the constraints on λ . Its informatics and software, as developed by Ali and Stutz of the Center for Cybernetic Studies of The University of Texas at Austin, involve only minor modification from that of Charnes, Cooper, Seiford and Stutz (1983).

5. Efficiency analysis

As mentioned, managerial and program comparison aspects of efficiency analysis were initiated in Charnes, Cooper and Rhodes (1978, 1981) and Charnes and Cooper (1980), through a generalization of the single-input, single-output absolute efficiency determination of classical engineering and science to multi-input, multi-output relative efficiencies of a finite number of decision-making units 'DMUs' (sometimes called 'productive' units or 'response' units). The multi-input, multi-output situations were reduced to 'virtual' single-input, single-output ones through use of virtual multipliers and

⁷See Charnes and Cooper (1961, ch. IX).

sums. Explicitly, the CCR ratio measure of efficiency of the DMU designated '0' is given by the non-linear, non-convex, non-Archimedean fractional program [see Charnes, Cooper, Lewin, Morey and Rousseau (forthcoming)]

$$\max_{\eta, \xi} \eta^T Y_0 / \xi^T X_0,$$

subject to

$$\eta^T Y_j / \xi^T X_j \leq 1, \quad -\eta^T / \xi^T X_0 \leq -\epsilon e^T, \quad -\xi^T / \xi^T X_0 \leq -\epsilon e^T, \quad (5.1)$$

for $j = 1, \dots, n$, where the entries of the X_j and Y_j are assumed positive, ϵ is a non-Archimedean infinitesimal, e^T is a row vector of ones, and, by abuse of notation, has s entries for η^T , m entries for ξ^T . (X_0, Y_0) is one of the n input-output pairs.

Employing the Charnes-Cooper transformation of fractional programming⁸

$$\mu^T \triangleq \eta^T / \xi^T X_0, \quad \nu^T \triangleq \xi^T / \xi^T X_0, \quad \nu^T X_0 = 1, \quad (5.2)$$

we obtain the dual non-Archimedean linear programs

$$\max_{\eta, \nu} \mu^T Y_0,$$

subject to

$$\nu^T X_0 = 1, \quad \mu^T Y - \nu^T X \leq 0, \quad -\mu^T \leq -\epsilon e^T, \quad -\nu^T \leq -\epsilon e^T, \quad (5.3)$$

and

$$\min_{\theta, \lambda, s^+, s^-} \theta - \epsilon e^T s^+ - \epsilon e^T s^-,$$

subject to

$$Y\lambda - s^+ = Y_0, \quad \theta X_0 - X\lambda - s^- = 0,$$

$$\lambda, s^+, s^- \geq 0,$$

where

$$X \triangleq [X_1, \dots, X_n], \quad Y \triangleq [Y_1, \dots, Y_n].$$

The first problem is associated with the origin of the term 'Data Envelopment Analysis' since the minimization (a) envelops the output vector Y_0 from above and (b) envelops the input vector X_0 from below via the minimizing choice of the scalar value of the intensity $\theta^* = \min \theta$. The second problem is said to be in efficiency analysis form with the maximization oriented toward the choice of μ^T and ν^T (called *virtual* multipliers or transformation rates)

⁸See Charnes and Cooper (1962) and Schaible (1974).

efficiency of the DMU designated θ -Archimedean fractional programming (Rousseau (forthcoming))

$$-\xi^T/\xi^T X_0 \leq -\epsilon e^T, \quad (5.1)$$

Y_j are assumed positive, ϵ is a vector of ones, and, by abuse of notation, (X_0, Y_0) is one of the n

of fractional programming⁸

$$1, \quad (5.2)$$

rams

$$r, \quad -v^T \leq -\epsilon e^T, \quad (5.3)$$

[the term 'Data Envelopment Analysis' is used below via the minimizing θ . The second problem is minimization oriented toward input or transformation rates)

which produces the greatest rate of virtual output per unit virtual input allowed by the first constraint together with the requirements (a) virtual output cannot exceed virtual input and (b) all virtual transformation rates must be positive.

Although, clearly, no assumptions have been made concerning the type of functional relations for the input-output pairs (X_j, Y_j) , the minimization program may be recognized as having the Data Envelopment Analysis constraints for an empirical production possibility set of Farrell, Shephard, etc. cone type $\mathcal{L}_E \cup \mathcal{B} \cup \mathcal{C}$, and, since

$$\theta - \epsilon[e^T Y \lambda - e^T X \lambda] \quad (5.4)$$

is an equivalent form for the functional, as being a Charnes-Cooper Pareto-optimality test for $(\theta X_0, Y_0)$ over the cone on the (X_j, Y_j) , $j = 1, \dots, n$, with pre-emption on the intensity θ of input X_0 . As mentioned above, DMU₀ is efficient iff $\theta^* = 1$, $s^{*-} = 0$, $s^{*+} = 0$.

Re informatics, which are particularly important since all n efficiency evaluations must be made (i.e., n linear programs must be solved), the dual problem can be computed exactly (in the base field) as shown in Charnes and Cooper (1961), e.g., with the code NONARC of Dr. I. Ali (Center for Cybernetic Studies, The University of Texas at Austin), or approximately by using a sufficiently small numerical value for ϵ . A typical efficient point is designated by $(\hat{x}, \hat{y})$ in fig. 1.

If a DMU is inefficient, the optimal $\lambda_j^* > 0$ in its DEA problem (= Charnes-Cooper test) designate efficient DMUs, as do alternate optima. Thus, a 'proper' subset of the efficient DMUs determines the efficiency value of an inefficient DMU. The convex combinations of this subset are also efficient. Thereby to each inefficient DMU a 'facet' of efficient DMU is associated. The transformation,

$$X_0 \rightarrow \theta^* X_0 - s^{*-}, \quad Y_0 \rightarrow Y_0 + s^{*+}, \quad (5.5)$$

where the asterisk designates optimality, projects DMU₀, i.e., (X_0, Y_0) , onto its efficiency facet.

This projection was introduced by Charnes, Cooper and Rhodes (1981) to correct for differences in managerial ability in order to distinguish between 'program' and 'managerial efficiency' in their analysis of programs Follow-Through and non-Follow-Through. It also shows quantitatively what improvements in inputs and outputs will (*ceteris paribus*) bring a DMU to efficient operation.⁹ Thus, although the relative efficiency measure of an inefficient

⁹The analysis by Bessent, Bessent, Charnes, Cooper and Thorogood (1983) shows how one might take account of the possible effects on other DMUs when one or more of the efficient DMUs is altered.

DMU will involve the infinitesimal ε , non-infinitesimal changes for improvement are suggested.

Both Farrell and Shephard knew that ratio measures required adjustments to correctly exhibit inefficiency of the second DMU in examples like the following two-input, one-output, two-DMU case:

DMU	x_1	x_2	y
1	1	2	1
2	1	4	1

Farrell added geometric points at infinity; Shephard simply excluded such cases without giving a method for their exclusion. The non-Archimedean extension in the CCR formulation was introduced to have an algebraically closed system of linear programming type. Linear programming theory holds for non-Archimedean as well as Archimedean entries in the vector and matrix problem data.¹⁰

Our new Pareto-efficient DEA method like in Charnes, Cooper, Seiford and Stutz (1983) associates facets with non-optimal (= non-Pareto-efficient) DMUs. Clearly, by the Charnes-Cooper test, DMU_0 is Pareto-efficient (Pareto-efficient) DMUs. Clearly, by the Charnes-Cooper test, DMU_0 is Pareto-efficient (Pareto-optimal) iff $-e^T s^{*+} - e^T s^{*-} = 0$, i.e., iff the l_1 -distance from (X_0, Y_0) to the farthest 'northwesterly' $\mathcal{Q}_e$ point is zero. The CCR efficient DMUs are also among the new Pareto-efficient DMUs. Projection of a non-optimal DMU onto its Pareto-efficient facet is rendered by

$$X_0 \rightarrow X_0 - s^{*-}, \quad Y_0 \rightarrow Y_0 + s^{*+}. \quad (5.6)$$

To achieve a convenient efficiency measure, we modify the functional by multiplying it by a $\delta > 0$ and dividing the s^+ and s^- by respectively the entries in Y_0 and X_0 , e.g.,

$$-\delta e^T D^{-1}(Y_0) s^+ - \delta e^T D^{-1}(X_0) s^-, \quad (5.7)$$

where $D^{-1}(Y_0)$ and $D^{-1}(X_0)$ are diagonal matrices whose diagonals are the reciprocals of the entries in Y_0 and X_0 respectively. This achieves a units-invariant measure which may be thought of as the *logarithm* of the efficiency measure. A $\delta = 10/(m+s)$ will yield a logarithm between 0 and -10 . This measure might then be called the 'efficiency pH' analogy with the pH of chemistry.

Our new measure relates to the units-invariant multiplicative measure of Charnes, Cooper, Seiford and Stutz (1983) which is necessary and sufficient

¹⁰See the discussion in Charnes and Cooper (1961, p. 756).

esimal changes for improve-
ures required adjustments to
examples like the following

ard simply excluded such
on. The non-Archimedean
d to have an algebraically
programming theory holds
es in the vector and matrix

arnes, Cooper, Seiford and
on-Pareto-efficient) DMUs.
Pareto-efficient (Pareto-effi-
, DMU₀ is Pareto-efficient
: l_1 -distance from (X_0, Y_0)
: CC efficient DMUs are
on of a non-optimal DMU

(5.6)

modify the functional by
by respectively the entries

(5.7)

whose diagonals are the
y. This achieves a units-
logarithm of the efficiency
etween 0 and -10. This
analogy with the pH of

multiplicative measure of
necessary and sufficient

that the DEA envelopments be piecewise Cobb-Douglas, by considering the entries in the X_j, Y_j to be logarithms of the entries in X_j, Y_j which we employ in the multiplicative formulation.

6. Informatics and function properties

6.1. Partial derivatives

The guidance provided by the Charnes-Cooper-Rhodes, Banker-Charnes-Cooper, Charnes-Cooper-Seiford-Stutz formulations does not include convenient access to the rates of change of the outputs with change in the inputs. The optimal dual variables in the DEA side linear programming problems give rates of change of the efficiency measure with changes in inputs or outputs. The non-Archimedean formulations further may give infinitesimal rates, which are not easily employed. And, for most of the efficient points one has non-differentiability because they are extreme points rather than (relative) interior points. Nevertheless, because of the informatics, e.g., computational tactics, we employ in testing via Charnes-Cooper for Pareto-efficiency, the following constructive method can be employed.

On reaching a non-Pareto-efficient point, our software discovers all the optimal observed points in its facet, hence, implicitly, all the convex combinations which form the facet. Since the Pareto-efficient facet is a linear surface it is not only differentiable everywhere in its relative interior but all its partial derivatives are constant throughout the facet. Thus, we need only obtain these for any relative interior point of the facet to have them for the whole facet.

Let

$$F(x_1, \dots, x_m, y_1, \dots, y_s) = 0 \quad (6.1)$$

be the linear equation of the facet. Since we have sufficient differentiability in the neighborhood of an interior point $(\bar{x}, \bar{y})$, we know

$$\partial y_i / \partial x_i |_{\bar{x}, \bar{y}} = -(\partial F / \partial x_i) / (\partial F / \partial y_i), \quad (6.2)$$

where the right-side partial derivatives are also evaluated at $(\bar{x}, \bar{y})$. Suppose we run the Charnes-Cooper test with $(\bar{x}, \bar{y})$ as the point being tested. Then the optimal dual variables corresponding to the input $\bar{x}_i$ and output $\bar{y}_i$ are respectively $(\partial F / \partial x_i) |_{\bar{x}, \bar{y}}$ and $(\partial F / \partial y_i) |_{\bar{x}, \bar{y}}$. Thus, the rate of change of output y_i with respect to input x_i is simply the negative of the ratio of the optimal dual x_i constraint variable to the optimal dual y_i constraint variable!¹¹

¹¹ See also Charnes, Cooper and Rhodes (1978, p. 439) for a discussion which can now relate this development to the ordinary conditions of economic theory for equality between ratios of marginal productivities and marginal rates of substitution.

More specifically, all Pareto-efficient (X_j, Y_j) of the facet for the point $(\bar{x}, \bar{y})$ satisfy

$$\mu^*{}^T y - \nu^*{}^T x - \varphi^* = 0, \quad (6.3)$$

where (μ^*, ν^*) are the dual evaluators at an optimal basic solution, since they do not depend on the Charnes-Cooper test right-hand sides. Thereby,

$$F(x, y) = \mu^*{}^T y - \nu^*{}^T x - \varphi^* = 0. \quad (6.4)$$

Clearly, $\mu_i^* = \partial F / \partial y_i$ and $-\nu_i^* = \partial F / \partial x_i$ as already stated.

It should be borne in mind, of course, that these rates of change are valid only for changes which keep one within the facet.

6.2. Isotonicity and economies of scale

To date the structure, or 'geometry', of empirical Pareto-efficient production functions has received little attention. The structure depends on and varies with the 'production possibility' or 'reference' set chosen. Here we make a beginning for our new set Q_E and leave to later research more in-depth and broader explorations including those for other reference sets.

In many practical situations we try to choose inputs and outputs with the thought that the underlying empirical Pareto-efficient function should be 'isotone' (which means 'order-preserving'). By definition a (vector) function $f(x)$ is isotone if $x^a \leq x^b$ implies $f(x^a) \leq f(x^b)$.¹² What in fact can be the case? We show here that in the single-output situation the empirical Pareto-efficient function is *always* isotone. The multiple-output situation, however, may only satisfy a weaker function property which we shall call 'c-d-isotone' or 'cone-directional-isotone', i.e., there is a cone of directions in output space on which the outputs projection is isotone.

Consider first the 'facets' of an empirical Pareto-efficient function for our reference set. These consist of convex combinations of the Pareto-efficient sample input-output points with respect to this reference set.

Theorem 3. *Every facet of the empirical Pareto-efficient function relation consists of Pareto-efficient input-output points.*

Proof. Each facet corresponds to a basic optimal solution to the following linear program (Charnes-Cooper test) for some sample input-output point

¹² The mathematical term 'isotone' is synonymous with the expression 'monotonically increasing'.

the facet for the point $(\bar{x}, \bar{y})$

(6.3)

ptimal basic solution, since
t-hand sides. Thereby,

(6.4)

dy stated.

ie rates of change are valid

Pareto-efficient production
re depends on and varies
chosen. Here we make a
search more in-depth and
ence sets.

puts and outputs with the
cient function should be
nition a (vector) function

What in fact can be the
the empirical Pareto-effi-
t situation, however, may
shall call 'c-d-isotone' or
ctions in output space on

efficient function for our
is of the Pareto-efficient
ence set.

ent function relation con-

olution to the following
iple input-output point

ssion 'monotonically increas-

(X_0, Y_0) :

$$\min_{\lambda, s^+, s^-} -e^T s^+ - e^T s^-,$$

subject to

$$Y\lambda - s^+ = Y_0, \quad -X\lambda - s^- = -X_0, \quad e^+ \lambda = 1,$$

$$\lambda, s^+, s^- \geq 0,$$

where $Y \triangleq (Y_1, \dots, Y_n)$, $X \triangleq (X_1, \dots, X_n)$, with optimal $s^{*+} = 0$, $s^{*-} = 0$. Every (X_j, Y_j) in the optimal basis is Pareto-efficient since (1) the optimal dual evaluators determined by the basic solution do not depend on the right-hand-side vector $(-X_0, Y_0)$, and (2), replacing $(-X_0, Y_0)$ by any $(-X_j, Y_j)$ in the basis preserves the feasibility of the basis for solution with the new right-hand side.

Next, if we Charnes-Cooper test any convex combination of the basic $(-X_j, Y_j)$, e.g., $(-X_B, Y_B)^T \theta_B$ with $\theta_B \geq 0$, $e_B^T \theta_B = 1$ for Pareto-efficiency and this reference set, i.e., insert it in place of $(-X_0, Y_0)^T$, an optimal λ^* is then simply θ_B (plus $s^{*+}, s^{*-} = 0$) with this same basis. Thus the whole facet is Pareto-efficient. ■

Now we prove:

Theorem 4. *If a Pareto-efficient empirical production function has only a single output, then it is an isotone function.*

Proof. Suppose $x^a \leq x^b$ for Pareto-efficient (x^a, y^a) and (x^b, y^b) . By definition of Pareto-efficiency for (x^b, y^b) , some component output y_j^a can exceed y_j^b only if some other output $y_k^a < y_k^b$. Since there is only one output, then $y^a \triangleq f(x^a) \leq y^b \triangleq f(x^b)$. Thus $x^a \leq x^b$ implies $f(x^a) \leq f(x^b)$ for Pareto-efficient points, i.e., the isotone property. ■

To show that a multiple-output Pareto-efficient production function need not be isotone, consider the following one-input, two-output example with three sample points a, b, c :

	a	b	c
x_1	1	2	3
y_1	5	2	1
y_2	5	7	4

Input-output points a and b are Pareto-efficient, c is not. We do *not* have isotonicity since $x^a < x^b$, but $f_1(x^a) \triangleq y_1^a > f_1(x^b) \triangleq y_1^b$.

If now we 'project' the outputs of a and b along the direction given by w_1 and w_2 for y_1 and y_2 we obtain the *single* output $5w_1 + 5w_2$ for a and $2w_1 + 7w_2$ for b . Requiring

$$7w_2 + 2w_1 \geq 5w_2 + 5w_1$$

implies $w_2 \geq \frac{3}{2}w_1$ as the cone of directions (w_1, w_2) which yields an isotonic relation.

In general when $x^c \geq x^d$, if $y_i^c \geq y_i^d$ for $i \in R^+$, where $R^+ \cup R = \{1, 2, \dots, s\}$, then the cone of isotonic directions $(w_1, \dots, w_s)$ is specified by

$$\sum_{i \in R^+} (y_i^c - y_i^d) \geq \sum_{i \in R} w_i (y_i^d - y_i^c). \quad (6.5)$$

Homogeneous production functions play an important role in the economics literature. Thereby, whether or not a function for which $f(\rho x) = \rho^\alpha f(x)$, with $\rho \geq 0$, had economies of scale would be decided by the value of the exponent α . More generally, increasing or decreasing 'returns to scale' would be present respectively, at $\bar{x}$ if $f(\rho \bar{x}) > \rho f(\bar{x})$ or $f(\rho \bar{x}) < \rho f(\bar{x})$ for $\rho > 1$ at points $\rho \bar{x}$ in a small neighborhood of $\bar{x}$. Banker, Charnes and Cooper (1984) give a criterion for deciding this (with production possibility set $\mathcal{Q}_E \cup \mathcal{B} \cup \mathcal{C}$ or $\mathcal{Q}_E \cup \mathcal{B}$), but does not give us the rates of change.

Because of our preceding theorems, however, we know that empirical Pareto-efficient functions are c-d-isotonic on facets and concave in each component function regardless of the nature of the underlying production possibility set. Thereby, we automatically anticipate lower and lower returns to scale in going from facet to facet with increasing $e^T x$. And our partial derivatives can give us explicitly the rates of change in each observed facet.

Practically, our choices of inputs are generally made with the expectation that the underlying Pareto-optimal function is isotonic, i.e., we choose the form of the inputs so that an increase in an input should not decrease the outputs. But even here we need still more to determine the non-concave portions of an isotonic function. For example, in fig. 2 an isotonic function is plotted together with the resulting concave cap (large dashed lines) obtained as the empirical function.

As suggested in our original paper,¹³ non-concavity can be explored by applying (output) component by component strictly concave transformations g_i to obtain $g_i(y_i)$ instead of y_i , so that $g_i(y_i(x))$ would be concave and out plot might look like fig. 3.

¹³'An Empirical DEA Production Function' by Charnes, Cooper and Seiford, April 1981, CCS 396, Center for Cybernetic Studies, The University of Texas, Austin, TX.

irical production functions

it, c is not. We do *not* have
 $y_1^b \triangleq y_1^h$.
 ing the direction given by w_1
 output $5w_1 + 5w_2$ for a and

z_2) which yields an isotonic

$r \in R^1$, where $R^1 \cup R^2 =$
 $w_1, \dots, w_n$) is specified by

(6.5)

ortant role in the economics
 which $f(\rho x) = \rho^\alpha f(x)$, with
 the value of the exponent
 to scale' would be present
 for $\rho > 1$ at points $\rho \bar{x}$ in a
 per (1984) give a criterion
 $\cup \mathcal{B} \cup \mathcal{C}$ or $\mathcal{Q}_E \cup \mathcal{B}$), but

we know that empirical
 its and concave in each
 ie underlying production
 ower and lower returns to
 g $e^T x$. And our partial
 in each observed facet.
 ade with the expectation
 z, i.e., we choose the form
 not decrease the outputs.
 n-concave portions of an
 action is plotted together
 btained as the empirical

ity can be explored by
 concave transformations
 uld be concave and out

and Seiford, April 1981, CCS
 , TX.

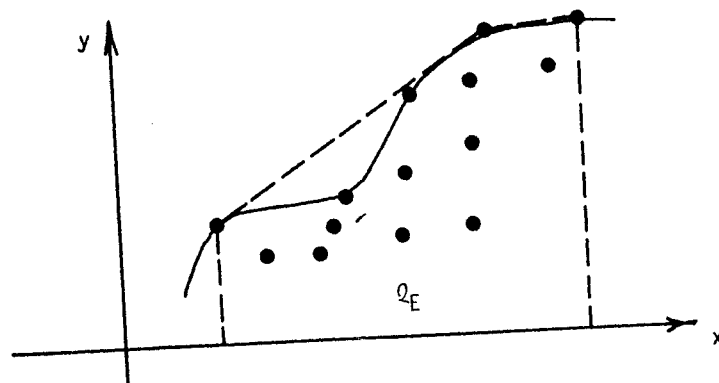


Fig. 2

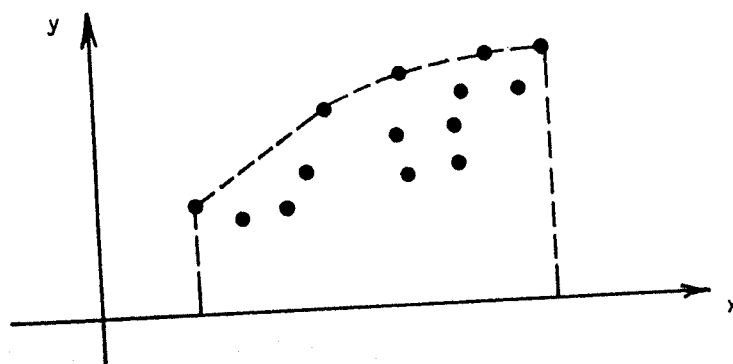


Fig. 3

Informatically, we can do this by applying transformations of form $g_i(y_i) = \bar{y}_i + (y_i - \bar{y}_i)^{1/\beta}$ with $\beta \approx 20$ to obtain possible new facets in the $g_i(y_i)$ (where $\bar{y}_i = \min_j y_{i,j}$).

Problems do arise, of course, on whether one gets spurious empirical frontier portions in this manner for empirical points which should 'really' be inefficient. Evidently such non-concave portions are portions of increasing returns to scale if they are truly on the frontier.

6.3. Discretionary and non-discretionary inputs

In a number of practical applications, certain relevant inputs (e.g., unemployment rate, population, median income) are not subject to 'discretionary'

change by the decision-makers of decision-making units. These are called 'non-discretionary' inputs.¹⁴ They are important in influencing the outputs and in furnishing the *reference* background in terms of which units' efficiency is rated. Not infrequently the facet associated with an inefficient unit has the same values for the non-discretionary inputs, in which case there is no problem with the rating assigned. If not, however, to obtain more meaningful ratings we can add constraints on λ to those in (4.5) which require the non-discretionary inputs to be the same as that of the unit being evaluated. Thereby, a more meaningful rating will be attained.

7. Conclusions

We have shown how direct application of the Charnes-Cooper test for Pareto-optimality leads to a simpler and more robust method, efficiency pH, encompassing all previous ones for ascertaining 'efficiency'. Further, Pareto-efficiency characterizations and constructions of empirical production functions restrict us methodologically to exploration of such functions by means of concave caps. Economies of scale from these thereby expectedly decrease with increase in the magnitude of the input vectors. Use of transformations of outputs, as we suggest, can uncover non-concave regions of the underlying production function where substantial economies of scale may prevail. Our new informatics device and theory of the use of the facet average (or bary-center) also constructively furnishes quantitative estimates of the rates of change of outputs with respect to inputs which have not been available previously. These new devices, as with other usages of empirical functions, suggest important new areas for development of statistical theory to distinguish between true properties and sampling 'accidents'. The vital importance of further development of the informatics of solution of systems of adaptively developed linear programming problems for Pareto-efficient constructions should also be clear.

References

- Afriat, S., 1972, Efficiency estimation of production functions, *International Economic Review* 13, 568-598.
- Aigner, D.J., C.A.K. Lovell and P.J. Schmidt, 1977, Formulation and estimation of stochastic frontier production function models, *Journal of Econometrics* 6, 21-37.
- Banker, R., 1984, Estimating most productive scale size using data envelopment analysis, *European Journal of Operational Research* 17, 35-44.
- Banker, R., A. Charnes and W.W. Cooper, 1984, Models for estimating technical and scale efficiencies in data envelopment analysis, *Management Science* 30, 1078-1092.

¹⁴See Charnes and Cooper (forthcoming).

g units. These are called
nfluencing the outputs and
' which units' efficiency is
an inefficient unit has the
h case there is no problem
ore meaningful ratings we
uire the non-discretionary
aluated. Thereby, a more

Charnes-Cooper test for
st method, efficiency p.H.,
ciency'. Further, Pareto-
pirical production func-
h functions by means of
expectedly decrease with
e of transformations of
gions of the underlying
scale may prevail. Our
facet average (or bary-
tima of the rates of
ave not been available
of empirical functions,
atistical theory to dis-
lents'. The vital impor-
olution of systems of
or Pareto-efficient con-

ational Economic Review 13,

and estimation of stochastic
11-37.
elopment analysis, European

imating technical and scale
1078-1092.

- Banker, R., W. Bowlin, A. Charnes and W.W. Cooper, in process, Mix inefficiencies and synergies in data envelopment analysis, CCS report (Center for Cybernetic Studies, The University of Texas, Austin, TX).
- Ben-Israel, A., A. Ben-Tal and A. Charnes, 1977, Necessary and sufficient conditions for a Pareto optimum in convex programming, *Econometrica* 45, 811-820.
- Bessent, A., W. Bessent, A. Charnes, W.W. Cooper and N. Thorogood, 1983, Evaluation of educational program proposals by means of DEA, *Educational Administration Quarterly* 19, 82-107.
- Charnes, A. and W.W. Cooper, 1962, Programming with linear fractional functionals, *Naval Research Logistics Quarterly* 9, nos. 3, 4.
- Charnes, A. and W.W. Cooper, 1977, Goal programming and multiple objective optimizations, *European Journal of Operational Research* 1, no. 1.
- Charnes, A. and W.W. Cooper, 1961, *Management models and industrial applications of linear programming* (Wiley, New York).
- Charnes, A. and W.W. Cooper, 1980, Auditing and accounting for program efficiency in not-for-profit entities, *Accounting, Organizations and Society* 5, 87-107.
- Charnes, A. and W.W. Cooper, forthcoming, Preface to topics in data envelopment analysis, *Annals of Operations Research*.
- Charnes, A., W.W. Cooper and E. Rhodes, 1978, Measuring the efficiency of decision-making units, *European Journal of Operations Research* 2, 429-444.
- Charnes, A., W.W. Cooper and E. Rhodes, 1981, Evaluating program and managerial efficiency: An application of data envelopment analysis to program follow through, *Management Science* 27, 668-687.
- Charnes, A., W.W. Cooper and A. Schinnar, 1982, Transforms and approximations in cost and production function relations, *Omega* 10, 207-211.
- Charnes, A., W.W. Cooper and D. Sherman, in process, A comparative study of DEA with ratio and regression approaches to efficiency evaluations, CCS report (Center for Cybernetic Studies, The University of Texas, Austin, TX).
- Charnes, A., W.W. Cooper, L. Seiford and J. Stutz, 1982, A multiplicative model for efficiency analysis, *Socio-Economic Planning Sciences* 6, 223-224.
- Charnes, A., W.W. Cooper, L. Seiford and J. Stutz, 1983, A units invariant multiplicative measure, *Operations Research Letters*.
- Charnes, A., W.W. Cooper, A. Lewin, R. Morey and J. Rousseau, forthcoming, An approach to positivity and stability analysis in DEA, *Annals of Operations Research*.
- Färe, R. and C.A.K. Lovell, 1978, Measuring the technical efficiency of production, *Journal of Economic Theory* 19, 150-162.
- Farrell, M.J., 1957, The measurement of productive efficiency, *Journal of Royal Statistical Society A* 120, 253-281.
- Fenchel, W., 1953, *Convex sets, cones and functions* (Princeton University Press, Princeton, NJ).
- Försund, F. and L. Hjalmarsson, 1979, Generalized Farrell measures of efficiency: An application to milk processing in Swedish dairy plants, *Economic Journal* 89, 294-315.
- Frisch, R., 1935, The principles of substitution: An example of its application in the chocolate industry, *Nordisk Tidsskrift for Teknisk Ekonomi* 1, 12-27.
- Rockafellar, T., 1970, *Convex analysis* (Princeton University Press, Princeton, NJ).
- Shephard, R.W., 1953, *Cost and production functions* (Princeton University Press, Princeton, NJ).
- Shephard, R.W., 1970, *Theory of cost and production functions* (Princeton University Press, Princeton, NJ).
- Schaible, S., 1974, Parameter-free convex equivalent and dual programs of fractional programming problems, *Zeitschrift für Operations Research* 18, 187-196.

SOME MODELS FOR ESTIMATING TECHNICAL AND SCALE INEFFICIENCIES IN DATA ENVELOPMENT ANALYSIS*

R. D. BANKER, A. CHARNES AND W. W. COOPER

School of Urban and Public Affairs, Carnegie-Mellon University,
Pittsburgh, Pennsylvania 15213

Graduate School of Business, University of Texas, Austin, Texas 78712

In management contexts, mathematical programming is usually used to evaluate a collection of possible alternative courses of action *en route* to selecting one which is best. In this capacity, mathematical programming serves as a planning aid to management. Data Envelopment Analysis reverses this role and employs mathematical programming to obtain *ex post facto* evaluations of the relative efficiency of management accomplishments, however they may have been planned or executed. Mathematical programming is thereby extended for use as a tool for *control* and evaluation of past accomplishments as well as a tool to aid in *planning* future activities. The CCR ratio form introduced by Charnes, Cooper and Rhodes, as part of their Data Envelopment Analysis approach, comprehends both technical and scale inefficiencies via the optimal value of the ratio form, as obtained directly from the data without requiring *a priori* specification of weights and/or explicit delineation of assumed functional forms of relations between inputs and outputs. A separation into technical and scale efficiencies is accomplished by the methods developed in this paper without altering the latter conditions for use of DEA directly on observational data. Technical inefficiencies are identified with failures to achieve best possible output levels and/or usage of excessive amounts of inputs. Methods for identifying and correcting the magnitudes of these inefficiencies, as supplied in prior work, are illustrated. In the present paper, a new separate variable is introduced which makes it possible to determine whether operations were conducted in regions of increasing, constant or decreasing returns to scale (in multiple input and multiple output situations). The results are discussed and related not only to classical (single output) economics but also to more modern versions of economics which are identified with "contestable market theories."

(EFFICIENCY; TECHNICAL INEFFICIENCY; RETURNS TO SCALE; MATHEMATICAL PROGRAMMING; LINEAR PROGRAMMING)

1. Background

Charnes, Cooper and Rhodes (CCR) (1978a, 1979) introduced a ratio definition of efficiency, also called the CCR ratio definition, which generalizes the single-output to single-input classical engineering-science ratio definition to multiple outputs and inputs without requiring preassigned weights. This is done via the extremal principle incorporated in the following model:

$$\max h_0 = \frac{\sum_{r=1}^s u_r y_{r0}}{\sum_{i=1}^m v_i x_{i0}} \quad \text{subject to} \quad 1 \geq \frac{\sum_{r=1}^s u_r y_{rj}}{\sum_{i=1}^m v_i x_{ij}}, \quad j = 1, \dots, n, \quad \text{with} \quad (1)$$

$$u_r, v_i > 0, \quad i = 1, \dots, m; \quad r = 1, \dots, s.$$

Here the $y_{rj}, x_{ij} > 0$ represent output and input data for decision making unit (DMU) j

*Accepted by Arie Y. Lewin; received December 11, 1981. This paper has been with the authors 4½ months for 2 revisions.

¹See Charnes, Cooper, Lewin, Morey, and Rousseau for an exact non-Archimedean expression of this "positivity" with necessary algebraic closure.

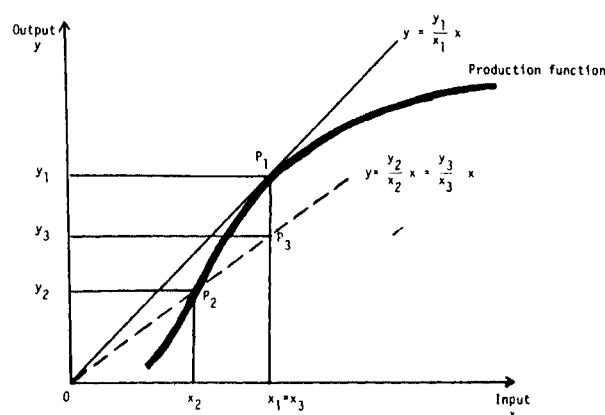


FIGURE 1

with the ranges for i , r and j indicated in (1). The data may be in the form of theoretically prescribed values or they may be in the form of observations. The unit to be rated is included in the functional with an index 0 as well as in the constraints, with the latter ensuring that an optimal $h_0^* = \max h_0$ will always satisfy $0 < h_0^* < 1$ with optimal solution values $u_r^*, v_i^* > 0$.

The main uses of these ideas have been in evaluations of "management" and "program" efficiencies² of decision making units (DMUs) of a not-for-profit variety such as schools,³ hospitals,⁴ etc. The ability to deal directly with multiple outputs and inputs forms one part of the appeal offered by these models and methods for uses such as these. Another part of its appeal comes from the development in Charnes, Cooper and Rhodes (1978a) which showed how the theory of fractional programming, as provided in Charnes and Cooper (1962), could be used to obtain access to a linear programming equivalent. This, in turn, yields an implementable form for securing solutions to (1) and it also yields a variety of duality relations for interpreting and utilizing the resulting $u_r^*, v_i^* > 0$.

Strong (and sharp) theoretical underpinnings as in physics and engineering are not available in applications such as we are considering. These must be replaced by weaker support—such as can be obtained from other disciplines like economics. It is, in fact, one purpose of the present paper to sharpen some of the latter contacts, but even after this has been accomplished, one must generally be satisfied with weaker results. For instance, one must be satisfied with a measure of only *relative* efficiency based on the available observations without recourse to what a stronger theory might provide.

We now try to clarify what has already been covered by reference to the illustration in Figure 1. Here we have portrayed the situation to be considered in terms of a single output, in amounts, y , and a single input, in amounts x . Three decision making units are to be rated for managerial efficiency. The production function represents the maximum output that can be produced for any specified input. The DMUs associated with P_2 and P_1 both achieve the maximum possible outputs for their input levels, while the DMU associated with P_3 falls short of the output level which is attainable from its x_3 input value.

To evaluate the efficiency of P_1 , we utilize (1), which, for this one output-one input

²See Charnes, Cooper and Rhodes (1981), for further discussion of differences in "management" and "program" efficiency.

³See Bessent, Bessent, Kennington and Reagan (1982).

⁴See D. Sherman (1982) which also contains an interesting comparison that highlights deficiencies of statistical regressions (including translog and Cobb-Douglas regressions) and econometric estimation and similar approaches that have been addressed to these multiple output situations in the past.

case, becomes:

$$\begin{aligned} \max h_0 = \frac{uy_1}{vx_1} \quad \text{subject to} \\ 1 \geq uy_1/vx_1, \quad 1 \geq uy_2/vx_2, \quad 1 \geq uy_3/vx_3, \quad u, v > 0, \end{aligned} \quad (2)$$

where x_i, y_i represent the input and output coordinates of the DMU associated with $P_i, i = 1, 2, 3$. The ray from the origin tangential to the production function at P_1 , lies above the ray through P_2 and P_3 . This means that the DMU associated with P_1 is efficient and the other two are not. In fact, we have

$$u^*y_1/v^*x_1 = 1 \quad \text{with also} \quad (3.1)$$

$$u^*y_2/v^*x_2 = u^*y_3/v^*x_3 < 1. \quad (3.2)$$

This result is readily verified for this very simple case in which we evidently have

$$y_1/x_1 > y_2/x_2 = y_3/x_3. \quad (3.3)$$

See Figure 1. Thus for all $u^*, v^* > 0$ we also have the equality portrayed in (3.2) and the other conditions in (3.1) and (3.2) are necessary if u^*, v^* are optimizing solutions to (2).

This results in a characterization of the DMUs associated with P_2 and P_3 as being "equally inefficient" relative to the DMU associated with P_1 . This characterization may be satisfactory in some cases. In other cases we may want to "fine tune" the developments in Charnes, Cooper and Rhodes (1978a) so that we can locate differences such as are portrayed in the P_2 and P_3 situations. Normally, of course, we will not have knowledge of the production functions but we can at least make a start toward this "fine tuning" to the extent that observational data may allow.⁵ It is toward this end that we shall direct our proposed contacts with economics even while recognizing that the concepts and definitions of theoretical economics as formulated for applications to private sector market behavior may not always be best suited for management science (and related) applications in the not-for-profit sectors.

2. Production Technologies and Efficiency Envelopes

The economic theory of production forms a natural point of contact with economics and we elect to take this route via the concepts of R. W. Shephard. To be sure, Shephard's work (1953, 1970) is primarily directed toward developing formal relations between cost functions and a corresponding production technology on the assumption that a theoretically known efficiency has already been attained.⁶ The production technology considered, however, encompasses the situation of multiple outputs in an unambiguous manner whereas other parts of (classical) production theory in economics are restricted to the single output situation—which is evidently of little or no interest in not-for-profit sector applications.⁷ In addition, Shephard has introduced the concept of a "distance function"⁸ which can be related to the important pioneering

⁵See Allen (1939) and Ferguson (1969) for treatments involving assumed knowledge of the production technologies. See also the survey by Kopp (1981).

⁶Note that economics concepts such as returns to scale, etc., have no unambiguous meaning until the efficiency frontier is attained. Thus, by virtue of this comment alone, most of the statistical-econometric studies on this topic are put in serious question. Other troubles may also be present as discussed in Charnes, Cooper and Schinnar (1982).

⁷Shortcomings of the classical single output theory even for purposes of economic theory are explained in Panzar and Willig (1977). See also Bailey and Friedlander (1982).

⁸Actually the distance function employed by Shephard is more properly regarded as a gauge function in the sense of Fenchel (1953).

For
each

and

V

F

scal

P

(t

P

P

Post

(X_j

⁹See

¹⁰Sh

efficien

¹¹In

$$u, v > 0, \quad (2)$$

the DMU associated with
production function at P_1 , lies
the DMU associated with P_1 is

$$(3.1)$$

$$(3.2)$$

which we evidently have

$$(3.3)$$

as portrayed in (3.2) and
are optimizing solutions

with P_2 and P_3 as being
the DMU. This characterization
is tantamount to "fine tune" the
model so that we can locate differ-
entially, of course, we will
be able to at least make a start
may allow.⁵ It is toward
economics even while
the model as formulated
ways be best suited for
profit sectors.

envelopes

in contact with economics
Shephard. To be sure,
developing formal relations
theory on the assumption
is needed.⁶ The production
multiple outputs in an
production theory in econo-
mics is evidently of little or no
importance and has introduced the
important pioneering

knowledge of the production

ambiguous meaning until the
of the statistical-econometric
model as discussed in Charnes,

economic theory are explained in

regarded as a gauge function in

work of M. J. Farrell⁹ in measuring efficiency directly from observational data, at least in the single output case. In this way, as shown in Banker (1980b), Shephard's distance function¹⁰ can be used to extend these ideas to more general situations, including those of a multiple output variety.

It should perhaps be explicitly observed that the CCR formulations in Charnes, Cooper and Rhodes (1981, 1978a) differ in important ways from the usual concepts of a production function at either the individual firm or aggregate level in both empirical and theoretical economics. In fact CCR refer to their function as an "envelope" developed relative to observational data from all of the $j = 1, \dots, n$ DMUs, with the envelope forming an efficiency frontier relative to each firm (= DMU) that is to be evaluated. It is to be borne in mind, however, that it is not always appropriate to regard this envelope as a production function in the usual (classical) senses for some of the uses to which the CCR formulations may be put.

The developments we shall use here will be via the kinds of axiomatic formulations which have become common in this part of economics as a result of Shephard's work (1953, 1970). We shall try to do this, however, in a way that maintains contact with the kind of analytic formulations that are required for implementing these ideas in efficiency evaluations to be obtained directly from observational data.

3. Axioms

Our approach will be via optimizations conducted with respect to already generated observations.¹¹ For this situation we shall construct a simple model with input-output configurations observed for each of $j = 1, \dots, n$ DMUs as (X_j, Y_j) , where $X_j = (x_{1j}, \dots, x_{mj})$ is a vector of observed inputs and $Y_j = (y_{1j}, \dots, y_{rj})$ is a vector of observed outputs for DMU j . It is assumed that at least one output and at least one input are positive. Every DMU j used for efficiency comparisons is assumed to have used the same inputs and produced the same outputs, although, in general, in varying amounts. Our objective is to characterize a production possibility set and, in particular, to determine an "efficient" subset based on these observed data.

We shall represent the production possibility set as

$$T = \{(X, Y) \mid Y \geq 0 \text{ can be produced from } X \geq 0\}. \quad (4)$$

Following Shephard (1970, p. 179) we then define the input possibility set $L(Y)$, for each Y , as

$$L(Y) = \{X \mid (X, Y) \in T\} \quad (5)$$

and the output possibility set $P(X)$, for each X , as

$$P(X) = \{Y \mid (X, Y) \in T\}. \quad (6)$$

We next postulate the following properties for the production possibility set, T :

Postulate 1. Convexity. If $(X_j, Y_j) \in T$, $j = 1, \dots, n$, and $\lambda_j \geq 0$ are nonnegative scalars such that $\sum_{j=1}^n \lambda_j = 1$, then $(\sum_{j=1}^n \lambda_j X_j, \sum_{j=1}^n \lambda_j Y_j) \in T$.

Postulate 2. Inefficiency Postulate. (a) If $(X, Y) \in T$ and $\bar{X} \geq X$, then $(\bar{X}, Y) \in T$.
(b) If $(X, Y) \in T$ and $\bar{Y} \leq Y$, then $(X, \bar{Y}) \in T$.

Postulate 3. Ray Unboundedness. If $(X, Y) \in T$ then $(kX, kY) \in T$ for any $k > 0$.

Postulate 4. Minimum Extrapolation. T is the intersection set of all $\hat{T}$ satisfying Postulates 1, 2 and 3, and subject to the condition that each of the observed vectors $(X_j, Y_j) \in \hat{T}$, $j = 1, \dots, n$.

⁹See Farrell (1957) and Farrell and Fieldhouse (1962).

¹⁰Shephard's distance function is also in need of fine tuning, as we shall see below, if it is to be used for efficiency characterizations. See also the discussion in Färe and Lovell (1978).

¹¹In this respect, we follow Farrell (1957).

Thus, T is the "smallest" set consistent with the observed data and the postulated properties for the production possibility set. Because T is based on convexification and ray extension, it is a polyhedral set.

Next we seek to characterize any $(X, Y) \in T$. Postulates 1 and 3 imply that every (X, Y) of the form $(k \sum_{j=1}^n \lambda_j X_j, k \sum_{j=1}^n \lambda_j Y_j)$ with $k > 0$, $\lambda_j \geq 0$ and $\sum_{j=1}^n \lambda_j = 1$ is in T . Furthermore, employing Postulates 2 and 4, we can deduce that

$$\begin{aligned} (X, Y) \in T \text{ if and only if } & X \geq k \sum_{j=1}^n \lambda_j X_j, \text{ and } Y \leq k \sum_{j=1}^n \lambda_j Y_j, \\ & \text{for some } k > 0 \text{ and some } \lambda_j, j = 1, \dots, n, \\ & \text{satisfying the conditions } \lambda_j \geq 0 \text{ and } \sum_{j=1}^n \lambda_j = 1. \end{aligned} \quad (7)$$

In the economics literature, Postulate 2 is sometimes referred to as "Free Disposability," a term which appears to have originated with Koopmans (1951). This may be a natural terminology for a discipline which has the analysis of market prices as a central concern. However, it risks a confounding of "market (allocative) efficiency" with the more basic condition of "technical efficiency" which is one of our central concerns.¹² Hence we prefer to use a term like "Efficiency Postulate" which we have converted to "Inefficiency Postulate" to indicate that inefficient production is always possible in the form of more inputs, smaller outputs, or both. In a similar way we prefer the more neutral term "ray unboundedness" to other alternatives such as "constant returns to scale." In this way we call attention to the fact that our usage of the latter is for individual characterizations and also to allow for other possibilities such as the presence of capacity limitations on inputs, especially when inefficiency possibilities are to be evaluated.

4. Shephard's Distance Function and Efficiency Measures

Having thus specified the production possibility set T , we next turn to the estimation of Shephard's distance function from these same observational data in order to relate it to the CCR efficiency measure. Shephard (1970, p. 206) defines a "distance function" $g(X, Y)$ of an input set $L(Y)$ as $g(X, Y) = 1/h(X, Y)$ where $h(X, Y) = \min\{h : hX \in L(Y), h > 0\}$.

Employing the characterization of any $(X, Y) \in T$ as in (4), we can express $h(X, Y)$ as:

$$\begin{aligned} h(X, Y) = \min h \quad \text{subject to} \quad & hX \geq k \sum_{j=1}^n \lambda_j X_j, \quad Y \leq k \sum_{j=1}^n \lambda_j Y_j, \\ & 1 = \sum_{j=1}^n \lambda_j, \quad \lambda_j \geq 0 \text{ and } k > 0. \end{aligned} \quad (8)$$

We next substitute $\mu_j = k\lambda_j$ in (8) above. Thus, we have

$$\begin{aligned} \min h \quad \text{subject to} \quad & hX - \sum_{j=1}^n \mu_j X_j \geq 0, \quad \sum_{j=1}^n \mu_j Y_j \geq Y, \\ & \mu_j \geq 0, \quad j = 1, 2, \dots, n. \end{aligned} \quad (9)$$

This is a linear programming problem, for which a dual can be written as

$$\begin{aligned} \max U^T Y \quad \text{subject to} \quad & V^T X = 1, \quad U^T Y_j - V^T X_j \leq 0, \quad j = 1, \dots, n, \\ & U \geq 0, \quad V \geq 0, \end{aligned} \quad (10)$$

where $U^T \equiv (u_1, \dots, u_r, \dots, u_s)$ and $V^T \equiv (v_1, \dots, v_i, \dots, v_m)$.

This is equivalent to the fractional programming problem with fractional con-

¹²The difference between these two types of efficiency are discussed in detail in Farrell (1957).

data and the postulated
on convexification and

and 3 imply that every
and $\sum_{j=1}^n \lambda_j = 1$ is in T .

$\sum_{j=1}^n \lambda_j X_j$,

(7)

as "Free Disposabil-
(1951). This may be a
of market prices as a
allocative) efficiency"
is one of our central
ulate" which we have
production is always
In a similar way we
alternatives such as
fact that our usage of
for other possibilities
ally when inefficiency

Measures

turn to the estimation
data order to relate
defines a "distance
 X, Y) where $h(X, Y)$

can express $h(X, Y)$

$$h(X, Y) = \frac{1}{k} \sum_{j=1}^n \lambda_j Y_j, \quad (8)$$

$$Y_j > Y, \quad (9)$$

can be written as

$$j = 1, \dots, n, \quad (10)$$

with fractional con-

in Farrell (1957).

straints:

$$\max h = \frac{U^T Y}{V^T X} \quad \text{subject to} \quad \frac{U^T Y_j}{V^T X_j} \leq 1, \quad j = 1, \dots, n, \quad U, V > 0. \quad (11)$$

This could describe the CCR efficiency measure, except that each u_r and v_i in U and V is only required to be nonnegative rather than strictly positive. See (1).

The positivity requirement would be obtained, however, if we followed the non-Archimedean formulation and development in Charnes, Cooper and Rhodes (1979),¹³ as we shall do below. Hence we shall say that we have established an equivalence between the CCR measure and the reciprocal of Shephard's distance function for input sets $L(Y)$, under the assumption that the production possibility set T satisfies the four postulates stated above.

Fine tuning is also required for Shephard's distance function, as already noted, if it is to be used for DMU efficiency measurement. Note, in particular, that the set $L_S(Y)$ defined by

$$L_S(Y) = \{X \mid h(X, Y) = 1\} \quad (12)$$

merely describes the boundary for inputs $X = (x_1, \dots, x_m)$ that can be used to produce a given combination of outputs $Y = (y_1, \dots, y_s)$. That is, following Shephard (1970, pp. 15-19) we allow for subsets of $L_S(Y)$ which we shall designate as

$$L_E(Y) = \{X \mid h(X, Y) = 1 \text{ with } \bar{X} < X \Rightarrow \bar{X} \notin L(Y) \text{ unless } \bar{X} = X\}. \quad (13)$$

This enables us to distinguish between points which are in the efficient subset $X_1 = (x_1, \dots, x_i, \dots, x_m) \in L_E(Y) \subseteq L_S(Y)$ and points which may be only on the boundary as in $X_2 = (x_1, \dots, x_i + d, \dots, x_m) \in L_S(Y)$ with $d > 0$.

The situation which enables us to distinguish between solutions like X_2 and X_1 in our linear programming formulation is the achievement of an optimum value of $h^* = 1$, but with positive slack in the i th input. In the situation for X_2 , efficiency is not achieved despite an $h^* = 1$ since this inefficient consumption for input i may be reduced to the level X_1 without affecting any other input or output.

We may clarify what is involved by introducing ideas from Charnes, Cooper and Rhodes (1978a). In particular, we replace (10) with the following problem:

$$\begin{aligned} \max z_0 = \sum_{r=1}^s u_r y_{r0} \quad \text{subject to} \quad \sum_{i=1}^m v_i x_{i0} = 1, \\ \sum_{r=1}^s u_r y_{rj} - \sum_{i=1}^m v_i x_{ij} < 0, \quad j = 1, \dots, n, \quad \text{and} \quad u_r > \epsilon, \quad v_i > \epsilon \quad \forall r, i, \end{aligned} \quad (14)$$

where $\epsilon > 0$ is a small "non-Archimedean" quantity.

The dual to this problem is

$$\begin{aligned} \min w_0 - \epsilon \left[\sum_{i=1}^m s_i + \sum_{r=1}^s s'_r \right] \quad \text{subject to} \\ 0 = w_0 x_{i0} - \sum_{j=1}^n x_{ij} \lambda_j - s_i, \quad i = 1, \dots, m, \end{aligned} \quad (15)$$

$$y_{r0} = \sum_{j=1}^n y_{rj} \lambda_j - s'_r, \quad r = 1, \dots, s, \quad \lambda_j, s_i, s'_r \geq 0, \quad \forall i, j, r.$$

Evidently we can then have $z_0^* = w_0^* = 1$ if and only if the slack variables s_i and s'_r are all zero at an optimum.

REMARK. Fare and Lovell (1978) discuss the relationship between Farrell's efficiency measure and Shephard's distance measure. The results are ambiguous, however, because they begin by defining Shep-

¹³See also Charnes, Cooper, Lewin, Morey and Rousseau for a further refinement which relates the slacks of the ratio problem to the slacks of the linear programming problem.

hard's distance measure to be the Farrell efficiency measure in a way that attributes properties of the Shephard distance measure to the Farrell measure as well, and an error in their "proof" (p. 154) leads them to conclude that the reciprocal of Shephard's output distance function (rather than the distance function itself) corresponds to Farrell's efficiency measure. They also fail to credit Farrell with the use of points at infinity, which avoids the problems they address. The above, following Charnes, Cooper and Rhodes (1978a), replaces this with a simple formulation in which the non-Archimedean elements provide what is required in the constraints of one problem and in the functional of the corresponding dual.

To complete the cycle of connections between CCR efficiency and Shephard's distance function, with the related non-Archimedean extensions, we next turn to the output sets $P(X)$ as defined in (6). For this set, Shephard (1970, p. 207) defines a distance function

$$h'(X, Y) = 1/g'(X, Y) \quad \text{where} \quad g'(X, Y) = \max\{g' \mid g'Y \in P(X), g' \geq 0\}. \quad (16)$$

By following an analysis similar to the one we have used for the distance function for input sets, it is easy to establish a similar relation between Shephard's distance function and the CCR measure (before adjustment for slacks) when T satisfies Postulates 1 to 4.

5. Reduction of Postulates

In order to extend what we have already achieved to the task of tracing the efficient production surface, we now delete Postulate 3 from our requirements. This "Ray Unboundedness" postulate enabled us to extrapolate the performance of the most efficient DMUs with efficient scale sizes (for their given input and output mixes) and identify any scale inefficiencies that may be reflected in the level of operations of other DMUs. By deleting this postulate we now restrict our attention strictly to production inefficiencies at the given level of operations for each DMU, and thus develop an efficiency measurement procedure that assigns an efficiency rating of one to a DMU if and only if the DMU lies on the efficient production surface, even when it may not be operating at the most efficient scale size. This identification of the efficient production surface will also allow us to determine whether increasing, constant or decreasing returns to scale¹⁴ prevail in different segments of the production surface. To reduce possible confusion with the concept of economies of scale, we shall avoid any recourse to cost functions¹⁵ and related price imputation requirements and deal directly with the inputs and outputs as they may be observed in vectors X_j, Y_j for the j th DMU.

Assuming now that the possibility set T satisfies only Postulates 1, 2, and 4, we proceed as before to characterize T as the "smallest" set satisfying the "convexity" and the "inefficiency" postulates, subject to the condition that each of the observed vectors $(X_j, Y_j) \in T$. Therefore, proceeding as in the previous section, we can deduce that a vector (X, Y) is in the set T if and only if

$$X \geq \sum_{j=1}^n \lambda_j X_j, \quad Y \leq \sum_{j=1}^n \lambda_j Y_j, \quad (17)$$

for some $\lambda_j \geq 0, j = 1, \dots, n$, satisfying the condition that $\sum_{j=1}^n \lambda_j = 1$.

We next determine Shephard's distance function for input sets $L(Y)$ when the production possibility set T is specified as above. Thus we have

$$g(X, Y) = 1/h(X, Y), \quad \text{where} \quad h(X, Y) = \min\{h \mid hX \in L(Y), h \geq 0\} \quad (18)$$

which we translate as

$$\begin{aligned} \min h = h(X_0, Y_0) \quad \text{subject to} \\ hX_0 - \sum_{j=1}^n \lambda_j X_j \geq 0, \quad \sum_{j=1}^n \lambda_j Y_j \geq Y_0, \quad \sum_{j=1}^n \lambda_j = 1, \quad \lambda_j \geq 0, \quad j = 1, \dots, n, \end{aligned} \quad (19)$$

¹⁴We are here referring to returns to scale in the restricted sense of theoretical economics. See §6 below.

¹⁵See, e.g., Farrell and Fieldhouse (1962) for further discussion.

it attributes properties of the "proof" (p. 154) leads them rather than the distance function derived from the use of points at Charnes, Cooper and Rhodes. These elements provide what is the corresponding dual.

inefficiency and Shephard's distance functions, we next turn to the definition of h (1970, p. 207) defines a

$$g' \in P(X), g' \geq 0. \quad (16)$$

the distance function for the j th DMU (when Shephard's distance function lacks) when T satisfies

task of tracing the efficient frontier. This "Ray" performance of the most efficient (input and output mixes) and relative of operations of other DMUs to production J , and thus develop an ordering of one to a DMU if even when it may not be the efficient production constant or decreasing return surface. To reduce shall avoid any recourse to and deal directly with Y_j for the j th DMU.

ulates 1, 2, and 4, we using the "convexity" and of the observed vectors, we can deduce that a

$$(17)$$

$$\sum_{j=1}^n \lambda_j = 1.$$

it sets $L(Y)$ when the

$$\in L(Y), h > 0 \} \quad (18)$$

$$(19)$$

$$j = 1, \dots, n,$$

cal economics. See §6 below.

relying on the fact that $h > 0$ will be satisfied when the components of every X_j and Y_j are all nonnegative—as is the case for the observational data we are considering.

This is a linear programming problem, the dual of which can be written as:

$$\max \sum_{r=1}^s u_r y_{r0} - u_0 \quad \text{subject to} \quad \sum_{r=1}^s u_r y_{rj} - \sum_{i=1}^m v_i x_{ij} - u_0 \leq 0, \quad j = 1, \dots, n, \quad (20)$$

$$\sum_{i=1}^m v_i x_{i0} = 1, \quad u_r, v_i \geq 0,$$

and u_0 is unconstrained in sign.

This linear programming problem is equivalent to a fractional programming problem,¹⁶ which we express as follows

$$\max \frac{\sum_{r=1}^s u_r y_{r0} - u_0}{\sum_{i=1}^m v_i x_{i0}} \quad \text{subject to} \quad \frac{\sum_{r=1}^s u_r y_{rj} - u_0}{\sum_{i=1}^m v_i x_{ij}} \leq 1, \quad \forall j, \quad u_r, v_i \geq 0, \quad (21)$$

with, again, u_0 unconstrained in sign.

Similarly, Shephard's distance function for output sets $P(X)$ can be expressed as:

$$\max h'(X, Y) = \frac{\sum_{r=1}^s u_r y_{r0}}{\sum_{i=1}^m v_i x_{i0} + v_0} \quad \text{subject to}$$

$$\frac{\sum_{r=1}^s u_r y_{rj}}{\sum_{i=1}^m v_i x_{ij} + v_0} \leq 1, \quad j = 1, \dots, n, \quad u_r, v_i \geq 0, \quad (22)$$

and v_0 is unconstrained in sign.

Of course, in general, these two distance functions will not be identical. Furthermore, we observe (as in the preceding section) that the set $L_S(Y) = \{X | h(X, Y) = 1\}$, where $h(X, Y)$ is the reciprocal of Shephard's input distance function, describes only a (boundary) isoquant. This isoquant may not coincide with the efficient subset $L_E(Y) = \{X \in L(Y) | \bar{X} \leq X, \bar{X} \in L(Y) \Rightarrow \bar{X} = X\}$.

Following the analysis in the previous section, we introduce the infinitesimal non-Archimedean quantity¹⁷ $\epsilon > 0$ to replace (20) with

$$\max \sum_{r=1}^s u_r y_{r0} - u_0 \quad \text{subject to} \quad \sum_{i=1}^m v_i x_{i0} = 1, \quad (20A)$$

$$-\sum_{i=1}^m v_i x_{ij} + \sum_{r=1}^s u_r y_{rj} - u_0 \leq 0, \quad j = 1, \dots, n, \quad u_r, v_i \geq \epsilon \quad \forall r, i,$$

and u_0 is unconstrained in sign.

In a similar vein, the dual problem in (19) may be rewritten as

$$\min h - \epsilon \left[\sum_{i=1}^m s_i^+ + \sum_{r=1}^s s_r^- \right] \quad \text{subject to}$$

$$h x_{i0} - \sum_{j=1}^n x_{ij} \lambda_j - s_i^+ = 0, \quad i = 1, \dots, m,$$

$$\sum_{j=1}^n y_{rj} \lambda_j - s_r^- = y_{r0}, \quad r = 1, \dots, s, \quad (19A)$$

$$\sum_{j=1}^n \lambda_j = 1, \quad \lambda_j, s_i^+, s_r^- \geq 0.$$

¹⁶We omit the transformation of variables implicit in this, as given explicitly in Charnes, Cooper and Rhodes (1978a), and, by abuse of notation, employ the same letters for the variables in both problems.

¹⁷As introduced in Charnes, Cooper and Rhodes (1978a, 1979).

6. Returns to Scale Characterizations

We now continue to extend our contacts with theoretical economics by turning to the problem of returns to scale (increasing, decreasing, or constant) as found in that literature. Once again, however, we emphasize that we are proceeding directly from observational data in our fine tuning of the CCR developments found in Charnes, Cooper and Rhodes (1978a). Mix and scale variations are likely to occur together in such observational data and so safeguards are required, as in our minimum extrapolation postulate, insofar as we cannot effect a separation of scale variations as in the formal theory of economics.¹⁸ We shall examine returns to scale locally at a point, say (X_E, Y_E) , on the efficient production surface, and relate it to the sign of the intercept term u_0 in the fractional programming problem (21) for this purpose.

We begin by asserting that the hyperplane given by

$$\sum_{r=1}^s u_r^* y_r - \sum_{i=1}^m v_i^* x_i - u_0^* = 0, \quad (23)$$

where the y_r, x_i are now variables, is a supporting hyperplane for the production possibility set T . Here u_r^*, v_i^* , and u_0^* are values of u_r, v_i , and u_0 that maximize the objective function in the fractional programming problem in (21).

We can confirm the support property of this hyperplane as follows. From the constraints to the problem in (21) we obtain

$$\sum_{r=1}^s u_r^* y_{rj} - \sum_{i=1}^m v_i^* x_{ij} - u_0^* \leq 0 \quad \text{for } j = 1, \dots, n. \quad (24)$$

Therefore, for any $\lambda_j > 0, j = 1, \dots, n$, with $\sum_{j=1}^n \lambda_j = 1$, we have

$$\sum_{r=1}^s u_r^* \sum_{j=1}^n y_{rj} \lambda_j - \sum_{i=1}^m v_i^* \sum_{j=1}^n x_{ij} \lambda_j - u_0^* \leq 0. \quad (25)$$

Further, by virtue of the result in (17), we can express any point $(X, Y) \in T$ as

$$\left(\sum_{j=1}^n \lambda_j X_j, \sum_{j=1}^n \lambda_j Y_j \right), \quad \text{where } \sum_{j=1}^n \lambda_j = 1, \lambda_j > 0, \text{ all } j.$$

Therefore, we have

$$(X, Y) \in T \Rightarrow \sum_{r=1}^s u_r^* y_r - \sum_{i=1}^m v_i^* x_i - u_0^* \leq 0. \quad (26)$$

Also, since (X_E, Y_E) is efficient we have

$$\frac{U^{*T} Y_E - u_0^*}{V^{*T} X_E} = 1 \quad \text{or} \quad U^{*T} Y_E - V^{*T} X_E - u_0^* = 0. \quad (27)$$

Thus (26) and (27) together imply that $U^{*T} Y - V^{*T} X - u_0^* = 0$ is a supporting hyperplane for T at the point (X_E, Y_E) .

Evidently this hyperplane is unique if and only if the optimal solution U^*, V^*, u_0^* to the equivalent linear programming problem is unique and this is the case we shall examine first. By virtue of (26) we have $U^{*T} Y - V^{*T} X - u_0^* \leq 0$ for all $(X, Y) \in T$. Hence in the unique supporting hyperplane situation which we are examining, a point (X_D, Y_D) in the neighborhood of (X_E, Y_E) will lie in the production possibility set if and only if $U^{*T} Y_D - V^{*T} X_D - u_0^* \leq 0$.

¹⁸See the discussion in Gold (1981) concerning the impracticality of effecting such separations from observational data so that characteristically the problem is usually approached from the cost function side. This, however, introduces problems with prices and allocative efficiencies.

economics by turning to
 (constant) as found in that
 proceeding directly from
 results found in Charnes,
 likely to occur together in
 our minimum extrapolation
 scale variations as in the
 locally at a point, say
 the sign of the intercept
 response.

(23)

and for the production
 and u_0 that maximize the
 (21).

as follows. From the

(24)

we

(25)

not $(X, Y) \in T$ as

0, all j .

(26)

(27)

$u_0^* = 0$ is a supporting

solution U^*, V^*, u_0^* to
 is the case we shall
 $u_0^* \leq 0$ for all $(X, Y) \in T$.
 are examining, a point
 production possibility set if

finding such separations from
 from the cost function side.

To ascertain whether increasing, constant or decreasing returns to scale are present at (X_E, Y_E) we let $Z_\delta \equiv [(1 + \delta)X_E, (1 + \delta)Y_E]$ be a point in the neighborhood of (X_E, Y_E) by choosing δ to be suitably small. Then we can say that

(28a) Increasing returns to scale are present if and only if there exists a $\delta^* > 0$ such that (1) $Z_\delta \in T$ for $\delta^* > \delta > 0$ and (2) $Z_\delta \notin T$ for $-\delta^* < \delta < 0$.

(28b) Constant returns to scale are present if and only if there exists $\delta^* > 0$ such that (1) $Z_\delta \in T$ for all δ such that $|\delta| < \delta^*$, or (2) $Z_\delta \notin T$ for all δ such that $0 < |\delta| < \delta^*$.

(28c) Decreasing returns to scale are present if and only if there exists $\delta^* > 0$ such that (1) $Z_\delta \notin T$ for $\delta^* > \delta > 0$ and (2) $Z_\delta \in T$ for $-\delta^* < \delta \leq 0$.

Now,

$$\begin{aligned} U^{*T}(1 + \delta)Y_E - V^{*T}(1 + \delta)X_E - (1 + \delta)u_0 \\ = (1 + \delta)(U^{*T}Y_E - V^{*T}X_E - u_0^*) + \delta u_0^* = \delta u_0 \end{aligned}$$

since we have $U^{*T}Y_E - V^{*T}X_E - u_0^* = 0$. Therefore, $Z_\delta \equiv ((1 + \delta)X_E, (1 + \delta)Y_E) \in T$ if and only if $\delta u_0^* \leq 0$ since we are restricting our attention to the case of unique supporting hyperplanes at E for the polyhedral set T . Employing (28a), (28b) and (28c), it follows immediately that in the case when a unique supporting hyperplane passes through an efficient point (X_E, Y_E) , we have

(29a) Increasing returns to scale $\Leftrightarrow u_0^* < 0$,

(29b) Constant returns to scale $\Leftrightarrow u_0^* = 0$,

(29c) Decreasing returns to scale $\Leftrightarrow u_0^* > 0$.

In other words, whether increasing, constant or decreasing returns to scale are present at (X_E, Y_E) depends on whether u_0^* is less than, equal to, or greater than zero in (20) and (21).

Figure 2 provides a generic picture of the situation in which scale changes are to be explored in the neighborhood of E . In the situation portrayed in Figure 2 we have $u_0^* > 0$ for the intercept value associated with the tangent line at E . The same situation would obtain for the succeeding piecewise segment on the efficient production possibility frontier. For the preceding segment where A appears, however, we would have $u_0^* < 0$ so that increasing returns would then be present. And, of course, for the

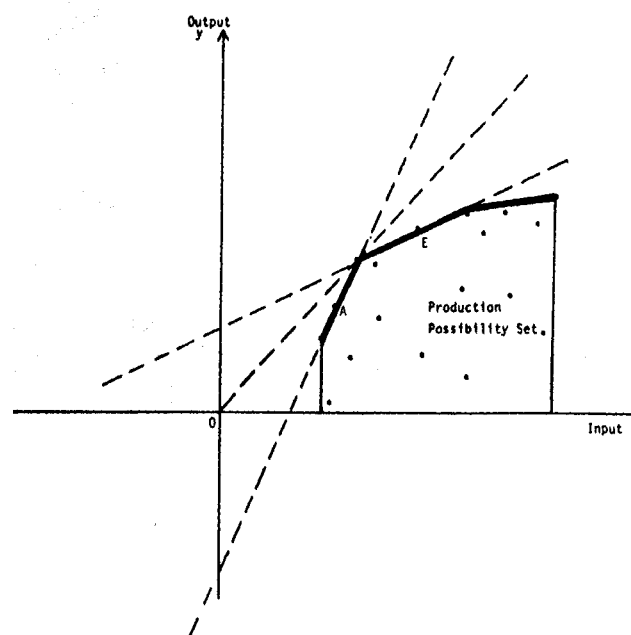


FIGURE 2

situation with $u_0^* = 0$ we would have constant returns if any such segment were present.

Up to this point we have been considering situations such as the one depicted generically at A and E in Figure 2 where the coefficients of the supporting hyperplane are determined by the components of the outward pointing normals at these points. At points of intersection such as shown for the two broken lines, these expressions are no longer unique. The components of the two normals at A and E provide alternate expressions, as do their convex combinations, from which additional supports may be designated that correspond to positions of rotation for these supporting hyperplanes at this point.

The above analysis considered efficiency measures based on input possibility sets as in (21). If we consider the alternative efficiency measures based on the output possibility sets, as in (22), then by pursuing a similar line of analysis we can relate the returns to scale at any point on the efficient production surface to the sign of the intercept term v_0^* for the supporting hyperplanes. In either case, in addition to providing a measure for the efficiency of individual observations and a means for tracing the efficient production surface, our mathematical programming formulation therefore also enables us to ascertain whether increasing, constant or decreasing returns to scale are present at specific points on the efficient production surface.

7. Concluding Interpretations

In this paper we have provided models for estimating technical and scale efficiencies of decision making units with reference to the efficient production frontier. The linear programming problems in (14) and (15) are employed to estimate the overall technical and scale efficiencies of a DMU. The linear programming formulations in (19A) and (20A) take into account the possibility that the average productivity at the most productive scale size may not be attainable for other scale sizes at which a particular DMU may be operating. These formulations estimate the pure technical efficiency of a DMU at the given scale of operation. The estimation of most productive scale size in DEA is discussed in Banker (1984).

Figure 3 illustrates these concepts of technical and scale efficiencies. The point A represents the DMU being evaluated. Its overall technical and scale efficiency is measured by the ratio MN/MA , by comparing the point A to the point N which reflects the average productivity attainable at the most productive scale size represented by the point E . The pure (input) technical efficiency of A is measured by the ratio MB/MA by comparing it with the point B on the efficient production frontier with the same scale size as A . Finally, the (input) scale efficiency of A is measured by the ratio MN/MB , so that the overall technical and scale efficiency MN/MA is equal to the product of the technical efficiency MB/MA and the scale efficiency MN/MB .

It is apparent from Figure 3 that the aggregate technical and scale efficiency measure MN/MA is less than the pure (input) technical efficiency measure MB/MA . This relationship between the two efficiency measures holds also for the general case of multiple inputs and outputs.

PROPOSITION. *The aggregate technical and scale efficiency measure as in (15) is less than or equal to the pure (input) technical efficiency measure as in (19A), with equality holding if and only if there exists an optimal solution to (15) such that the sum of the optimal values of the weights λ_j^* , $j = 1, \dots, n$, adds up to one. The proof follows immediately by comparing the constraint sets for the two formulations in (15) and (19A).*

These developments suggest that a scale efficiency measure can be defined as the ratio of the aggregate efficiency measure from (15) to the technical efficiency measure

with $S > 1$ if and only if increasing returns to scale²⁰ are present at (x, y) . Here $\phi(X, Y) \geq 0$ is a known transformation function which is assumed to be differentiable at every point (X, Y) satisfying $\phi(X, Y) = 0$ where X and Y represent input and output vectors, respectively. Since attention is restricted to one firm at a time in Panzar and Willig (1977), no additional identifying subscript is needed. Finally, it is assumed that each such firm conducts all of its operations on the efficiency frontier.²¹ How this is all to be implemented, tested or validated is not discussed in Panzar and Willig.

In our (DEA) approach we do not require knowledge of the transformation function and we also do not assume that each firm will attain the efficiency frontier. Instead, we employ the adjustment procedures described in Charnes, Cooper and Rhodes (1978a) in order to obtain new values $\hat{x}_{ij}, \hat{y}_{ij}$ which are all on the (relative) efficiency frontier for each of the $j = 1, \dots, n$ "firms" (= DMUs) represented in the observations. With these thus adjusted input and output values we will then have

$$1 = \frac{\sum_{r=1}^s u_r^* \hat{y}_{rj} - u_0^*}{\sum_{i=1}^m v_i^* \hat{x}_{ij}} \quad (31)$$

where we have indexed u_0^* , as from, e.g., repeated applications of (21), to provide the relevant identification for each of these $j = 1, \dots, n$ firms.

This can be transformed straightforwardly into

$$\frac{\sum_{i=1}^m v_i^* \hat{x}_{ij}}{\sum_{r=1}^s u_r^* \hat{y}_{rj}} = 1 - \frac{u_0^*}{\sum_{r=1}^s u_r^* \hat{y}_{rj}} = \hat{S}, \quad (32)$$

with $\hat{S} > 1$ if and only if $u_0^* < 0$, i.e., if and only if increasing returns to scale are present, and $\hat{S} \leq 1$ otherwise. This then is our analogue for the result obtained by Panzar and Willig.

Note that we have not required their assumptions of differentiability and knowledge of the transformation function. Of course, if such knowledge is available then the measure of efficiency and adjustment to the frontier defined by that function can be effected in a straightforward manner.²² In other cases, however, the DEA approach permits both efficiency measurement and adjustment to the relative frontier in a readily implemented manner.²³

As Panzar and Willig note, their results apply only locally, and the same is true for $\hat{S}$. Here again, however, an extension has been effected and given operational form. Thus whereas the results in Panzar and Willig (1977) apply only at the point (X, Y) , our results apply over the entire facet formed from the set of basis vectors used to evaluate DMU_{*j*}.²⁴

Reference to (32) also provides insight into paths of possible further extension in these DEA approaches. Comparing the terms on the far left and far right, it becomes clear that our separation into scale and technical efficiency is obtained by relaxing the "unity-or-less" condition on the original CCR ratio formulation as given in (1) on p. 1078, above.

²⁰ Another approach to returns to scale which is suited to DEA and does not rely on the Panzar and Willig assumptions may be found in Banker (1984).

²¹ It is assumed that the transformation function is also optimal since otherwise this firm may have its markets contested by new entrants. This all forms part of what is called "contestable market theory." As summarized in Baumol (1982), this theory retains the behavioral (optimizing) assumptions of the classical theory of competitive markets but (a) omits the requirement of having a large number of small firms and (b) switches the focus from single output to multiple output production.

²² See the discussion in Charnes and Cooper (1980).

²³ A code for doing this and other parts of DEA which has been developed by I. Ali and J. Stutz is available from the Center for Cybernetic Studies, BEB200E, The University of Texas at Austin, Texas 78712.

²⁴ The computer code described in the preceding footnote identifies the facet and the basis from which it is formed.

be present at (x, y) . Here y is assumed to be differentiable and represent input and output at a time in Panzar and Willig. Finally, it is assumed that y is on the frontier.²¹ How this is all done is given in Panzar and Willig.

The transformation function $f(x)$ is the efficiency frontier. Instead, we use the Dooper and Rhodes (1978a) relative efficiency frontier in the observations. With this we have

$$(31)$$

using (21), to provide the

$$(32)$$

assuming returns to scale are constant or the result obtained by

differentiability and knowledge of the transformation function available then the DEA approach to the relative frontier in a

and the same is true for a given operational form. Only at the point (X, Y) , the basis vectors used to

possible further extension in the and far right, it becomes a result obtained by relaxing the transformation as given in (1) on p.

do not rely on the Panzar and Willig

otherwise this firm may have its contestable market theory." As long as the assumptions of the classical large number of small firms and

developed by I. Ali and J. Stutz is at the University of Texas at Austin, Texas

and the basis from which it is

Instead of pursuing the latter course,²⁵ we now close by simplifying the above analysis in a way that makes contact with classical, single output, production theory in economics. This can be done by first replacing the above transformation function with a production function $f(x)$ which is applicable in this single output case.

Removing the identifying subscript j , we write $y = f(x)$ for the output which is (uniquely) obtained from input x . Then we adjust the denominator in (32) for this single (efficient) output situation and obtain

$$\hat{S} = 1 - \frac{u_0^*}{y} = \frac{y - u_0^*}{y}. \quad (33)$$

If we further assume that x represents a single input then we can also write

$$y = x \frac{\partial f}{\partial x} + u_0^* \quad (34)$$

where $\partial f / \partial x$ and u_0^* represent the slope and intercept values associated with, e.g., the pertinent linear segment in Figure 3.

Substituting from (34) into the numerator on the right-hand side of (33) gives

$$\hat{S} = x \frac{\partial f}{\partial x} / y = \frac{\partial f}{\partial x} / \frac{y}{x}. \quad (35)$$

Thus we have increasing, constant or decreasing returns to scale according to whether

$$\frac{\partial f}{\partial x} / \frac{y}{x} \gtrless 1 \quad \text{or equivalently,} \quad (36)$$

$$\partial f / \partial x \gtrless y / x. \quad (37)$$

This is the well-known result in classical economics²⁶ that returns to scale are increasing, constant or decreasing at any pair of coordinate values (x, y) according to whether marginal product is greater than, equal to, or less than average product. In other words, all aspects of that theory are preserved when it is applicable. In other cases we may use DEA to augment or replace it without breaking contact with the underlying concepts.²⁷

²⁵One such possibility involves the potentially important "Congestion of Production Factors" introduced in Färe and Svensson (1980) as a possible source of inefficiency.

²⁶See Allen (1939).

²⁷The situation is not unlike the actual use of dual variables in linear programming in place of the *ceteris paribus* assumptions of classical economics which mainly serve as only guides to the concept of marginal opportunity cost because these conditions can only rarely be realized.

Parts of this paper rest on developments in R. D. Banker's doctoral thesis, "Studies in Cost Allocation and Efficiency Evaluation," at the Graduate School of Business of Harvard University, Boston, Massachusetts 02163, 1980. This research was supported in part by Project NR 1947-021, ONR Contract N00014-81-C-0236 with the Center for Cybernetic Studies, The University of Texas and ONR Contract N00014-81-C-0410 at Carnegie-Mellon University's School of Urban and Public Affairs. Reproduction in whole or in part is permitted for any purpose of the United States Government.

References

- ALLEN, R. G. D., *Mathematical Analysis for Economists*, The MacMillan Co., New York, 1939.
- BAILEY, E. E. AND A. F. FRIEDLANDER, "Market Structure and Multiproduct Industries," *J. Econom. Lit.*, 20 (September 1982), 1024-1048.
- BANKER, R. D., "A Game Theoretic Approach to Measuring Efficiency," *European J. Oper. Res.*, (1980a), 262-266.
- , "Studies in Cost Allocation and Efficiency Evaluation," unpublished doctoral thesis, Harvard University, Graduate School of Business Administration, 1980. Also available from University Microfilms International, Ann Arbor, Mich., 1981.

- , "Estimating Most Productive Scale Size Using Data Envelopment Analysis," *European J. Oper. Res.*, (1984).
- , A. CHARNES, W. W. COOPER AND A. SCHINNAR, "A Bi-Extremal Principle for Frontier Estimation and Efficiency Evaluation," *Management Sci.*, 27 (December 1981).
- , R. CONRAD AND R. P. STRAUSS, "An Application of DEA to the Empirical Investigation of a Hospital Production Function," *Management Sci.* (submitted).
- BAUMOL, W., "Contestable Markets: An Uprising in the Theory of Industry," *Amer. Econom. Rev.*, 72, 1 (March 1982), 1-15.
- BESSENT, A., W. BESSENT, J. KENNINGTON AND B. REAGAN, "An Application of Mathematical Programming to Assess Productivity in the Houston Independent School District," *Management Sci.* 28, 12 (December 1982).
- CHARNES, A. AND W. W. COOPER, "An Efficiency Opening for Managerial Accounting in Not-For-Profit Entities," in P. Holzer (Ed.), *Proceedings of a Conference on Managerial Accounting*, University of Illinois Department of Accounting, Urbana, 1980.
- AND ———, "Managerial Economics: Past, Present and Future," *J. Enterprise Management*, 1 (1978).
- AND ———, "Programming with Linear Fractional Functionals," *Naval Res. Logist. Quart.*, 9, 3, 4 (1962).
- , A. LEWIN, R. MOREY AND J. ROUSSEAU, "An Approach to Positivity and Stability Analysis in DEA," *Ann. Oper. Res.* (submitted).
- , AND E. RHODES, "Data Envelopment Analysis: Approach for Evaluating Program and Managerial Efficiency with an Application to the Program Follow Through Experiment in U.S. Public School Education," *Management Sci.*, 27, 6 (June 1981), 668-697.
- , AND ———, "Measuring the Efficiency of Decision Making Units," *European Journal of Operational Research*, (1978a), 429-444.
- , AND ———, "Short Communication: Measuring the Efficiency of Decision Making Units," *European J. Oper. Res.*, (1979), 339.
- , AND A. SCHINNAR, "Transforms and Approximations in Cost and Production Function Relations," *Omega*, 10, 2 (1982), 207-211. See also accompanying "Prologue."
- , AND L. SEIFORD, "An Empirical DEA Production Function," 1981. Reissued as CCS Report 459 "Pareto-Optimality, Efficiency Analysis and Empirical Production Functions," by A. Charnes, W. W. Cooper, B. Golany, L. Seiford and J. Stutz, University of Texas at Austin, 1983.
- FÄRE, R. AND C. A. K. LOVELL, "Measuring the Technical Efficiency of Production," *J. Econom. Theory*, (1978), 150-162.
- AND L. SVENSSON, "Congestion of Production Factors," *Econometrica*, (November 1980), 1745-1753.
- FARRELL, M. J., "The Measurement of Productive Efficiency," *J. Roy. Statist. Soc. Ser. A*, 120 (1957), 253-290.
- AND M. FIELDHOUSE, "Estimating Efficient Production Frontiers under Increasing Returns to Scale," *J. Roy. Statist. Soc. Ser. A*, (1962), 252-267.
- FENCHEL, W., *Convex Sets, Cones and Functions*, Princeton University, Department of Mathematics, Princeton, N.J., 1953.
- FERGUSON, C. E., *The Neoclassical Theory of Production and Distribution*, Cambridge University Press, Cambridge, 1969.
- FÖRSUND, F. R., C. A. K. LOVELL AND P. SCHMIDT, "A Survey of Frontier Production Functions and of their Relationship to Efficiency Measurement," *J. Econometrics*, (1980).
- GOLD, B., "Changing Perspectives on Size, Scale and Returns: An Interpretive Survey," *J. Econom. Lit.*, (March 1981).
- KOOPMANS, T., *Activity Analysis of Production and Allocation*, John Wiley & Sons, Inc., New York, 1951.
- KOPP, R. J., "The Measurement of Productive Efficiency: A Reconsideration," *Quart. J. Econom.*, (1981), 477-503.
- PANZAR, J. C. AND R. D. WILLIG, "Economies of Scale in Multi-Output Production," *Quart. J. Econom.*, (August 1977), 481-493.
- RHODES, E., "Data Envelopment Analysis and Related Approaches for Measuring the Efficiency of Decision Making Units with an Application to Program Follow Through in U.S. Public School Education," Ph.D. Thesis, Carnegie-Mellon University, School of Urban and Public Affairs, Pittsburgh, 1978. Also available from University Microfilms International, Ann Arbor, Mich., 1979.
- SHEPHARD, R. W., *Cost and Production Functions*, Princeton University Press, Princeton, N.J., 1953.
- , *The Theory of Cost and Production Functions*, Princeton University Press, Princeton, N.J., 1970.
- SHERMAN, D., "Data Envelopment Analysis and Other Approaches for Measuring the Efficiency of Hospital Operations," DBA Thesis, Harvard University Graduate School of Business, 1981. Also available from University Microfilms International, Ann Arbor, Mich., 1982.

APPENDIX B

1. International Journal of Systems Science Article by Charnes, Cooper, Seiford and Stutz.
2. Operations Research Letters Article by Charnes, Cooper, Seiford and Stutz.

extra copies

A MULTIPLICATIVE MODEL FOR EFFICIENCY ANALYSIS†

A. CHARNES, W. W. COOPER and L. SEIFORD

Center for Cybernetic Studies, Business-Economics Building, The University of Texas at Austin, Austin, TX 78712, U.S.A.

and

J. STUTZ

University of Miami, Department of Management Science and Computer Information Systems, Coral Gables, FL 33124, U.S.A.

(Received 24 November 1981)

Abstract—This paper develops theory and algorithms for a "multiplicative" Data Envelopment Analysis (DEA) model employing virtual outputs and inputs as does the CCR ratio method for efficiency analysis. The frontier production function results here are of piecewise log-linear rather than piecewise linear form.

INTRODUCTION

In [1-3] Charnes, Cooper and Rhodes presented a method, the CCR ratio, for measuring the relative efficiency of DMU's (Decision Making units) having the same multiple inputs and multiple outputs. The essential characteristic of the CCR construction was the reduction of the multiple output-multiple input DMU situation to that of single virtual outputs and virtual inputs. For these, the ratios of virtual outputs to virtual inputs could be used to define relative efficiency in a manner similar to that in engineering practice.

Although it was understood from the first that the "additive" combinations of outputs or inputs to achieve virtual outputs and inputs was not the only way of obtaining virtual single outputs and inputs, to date no other way has been studied. It is the purpose of this paper to present a "multiplicative" combinational method. As will be seen, it has a theory similar to that of the CCR ratio and its Data Envelopment Analysis but one which is simpler in certain respects, which can prove advantageous in applications, interpretations and extensions of other theories.

A MULTIPLICATIVE EFFICIENCY MODEL

As in [1], suppose there are n DMU's and the j th DMU has output vector Y_j with s outputs and input vector X_j with m inputs. Suppose that our measures of inputs and outputs are such that all in our "sample" on n DMU's have values greater than unity.

To determine the relative efficiency of DMU₀ with input-output pair (X_0, Y_0) , we form the "multiplicative" virtual outputs and inputs for the DMU's as follows:

$$\prod_{r=1}^s Y_{rj}^{\mu_r} \text{ and } \prod_{i=1}^m X_{ij}^{\nu_i}, j = 1, \dots, n. \quad (1)$$

This construction is analogous to a geometric mean construction in that we take the exponents μ_r and ν_i to be non-negative, and in order that they all be positive (to avoid Pareto inefficiencies) we require them to be not less than unity.

The relative efficiency of DMU₀ will then be given as the maximum of the ratio of its virtual output to virtual input subject to the condition that the virtual output to virtual input ratio of *all* DMU's be less than or equal to unity i.e. we solve the problem

$$\begin{aligned} \max_{\mu, \nu} & \quad \prod_{r=1}^s Y_{r0}^{\mu_r} / \prod_{i=1}^m X_{i0}^{\nu_i} \\ \text{s.t.} & \quad \prod_{r=1}^s Y_{rj}^{\mu_r} / \prod_{i=1}^m X_{ij}^{\nu_i} \leq 1, j = 1, \dots, n \\ & \quad \mu_r \geq 1, r = 1, \dots, s \\ & \quad \nu_i \geq 1, i = 1, \dots, m. \end{aligned} \quad (2)$$

Taking logarithms (to any base), this may be written as the linear programming problem

$$\begin{aligned} \max & \quad \sum_{r=1}^s \mu_r \hat{Y}_{r0} - \sum_{i=1}^m \nu_i \hat{X}_{i0} \\ \text{s.t.} & \quad \sum_{r=1}^s \mu_r \hat{Y}_{rj} - \sum_{i=1}^m \nu_i \hat{X}_{ij} \leq 0, j = 1, \dots, n \\ & \quad -\mu_r \leq -1, r = 1, \dots, s \\ & \quad -\nu_i \leq -1, i = 1, \dots, m \end{aligned} \quad (3)$$

where the caret sign ($\hat{\cdot}$) denotes logarithms. More compactly this can be written as

$$\begin{aligned} \max & \quad \mu^T \hat{Y}_0 - \nu^T \hat{X}_0 \\ \text{s.t.} & \quad \mu^T \hat{Y} - \nu^T \hat{X} \leq 0 \\ & \quad -\mu^T \leq -e^T \\ & \quad -\nu^T \leq -e^T \end{aligned} \quad (4.1)$$

†This research is partly supported by ONR contract N00014-81-C-0236 and USARI contract MD-903-81-C-0365 with the Center for Cybernetic Studies, The University of Texas at Austin. Reproduction in whole or in part is permitted for any purpose of the United States Government.

with dual

$$\begin{aligned} \min \quad & -e^T s^+ - e^T s^- \\ \text{s.t.} \quad & \bar{Y}\lambda - s^+ = \bar{Y}_0 \\ & -\bar{X}\lambda - s^- = -\bar{X}_0 \\ & \lambda, s^+, s^- \geq 0 \end{aligned} \quad (4.2)$$

where $\bar{Y} = [\bar{Y}_1, \dots, \bar{Y}_n]$, $\bar{X} = [\bar{X}_1, \dots, \bar{X}_n]$ and the e^T are row vectors of ones of the appropriate sizes.

Evidently (4.2) is a DEA (Data Envelopment Analysis) problem with the cone on the DMU vectors enveloping $\bar{Y}_0$ from above, e.g. $\bar{Y}\lambda \geq \bar{Y}_0$ and enveloping $\bar{X}_0$ from below e.g. $\bar{X}\lambda \leq \bar{X}_0$. It is a simpler envelopment than that for the CCR ratio model. Further, its log-efficiency is given solely in terms of the optimal slacks in (4.2), whereas the CCR efficiency value also involves a change in intensity of the X_0 vector.

CHARACTERIZING EFFICIENT DMU'S

Since the optimal functional value in (4.2) is the log-efficiency of DMU_0 and a DMU is efficient if and only if it has a log-efficiency of zero, we have the result,

Theorem 1. DMU_0 is efficient in an optimal solution of (4.2) iff all slacks are zero.

Again, as in the CCR ratio model, the efficiency rating of a DMU is characterized by a certain subset of efficient DMU's. For,

Theorem 2. If DMU_k appears in an optimal basic solution to (4.2), then DMU_k is efficient.

Proof. The log-efficiency problem of DMU_k replaces the right hand side in (4.2) by $(\bar{Y}_k, -\bar{X}_k)^T$. An optimal basic solution for DMU_0 which contains the basis vector $(\bar{Y}_k, -\bar{X}_k)^T$ is feasible for the DMU_k problem with $\lambda_k = 1$, $\lambda_j = 0$, $j \neq k$ and $s^+, s^- = 0$. It is also optimal since the dual evaluators do not involve the right hand side. Thus, DMU_k has log-efficiency zero.

Also, as in the CCR ratio model, we can project a DMU onto the efficiency facet which determines its efficiency, for

Theorem 3. If the coefficient vectors of s_h^+, s_h^- are basic slack vectors in an optimal basic solution of (4.2) for DMU_0 , then replacing $(\bar{Y}_0, -\bar{X}_0)^T$ by $(\bar{Y}_0 + s_h^+, -\bar{X}_0 + s_h^-)^T$ yields an efficient DMU.

Proof. If $s^{*+}, s^{*-} = 0$, DMU_0 is efficient. If $e^T s^{*+} + e^T s^{*-} > 0$, then DMU_0 is not efficient. By Theorem 2, its output-input vector cannot be part of this optimal basis for it. Evidently the vector $(\bar{Y}_0 + s_h^+, -\bar{X}_0 + s_h^-)^T$ as a right hand side vector is feasible for this basis and with zero slacks. Thus, a fortiori, this basic solution is also a basic solution with both the right hand side and left hand side $(\bar{Y}_0, -\bar{X}_0)^T$ replaced by the new vector.

To confirm that this basic solution is still optimal for the new problem, we need only verify that the dual evaluators (μ^*, ν^*) satisfy the new constraint in the dual problem. Now our old optimal solution is the (unique) solution to

$$\begin{aligned} \bar{Y}_B \lambda_B^* - s_h^{*+} &= \bar{Y}_0 \\ -\bar{X}_B \lambda_B^* - s_h^{*-} &= -\bar{X}_0 \end{aligned} \quad (5)$$

where the "B" subscript designates the column vectors

corresponding to this basic solution. Thereby

$$\begin{aligned} \mu^{*T}(\bar{Y}_0 + s_h^+) &= \mu^{*T}(\bar{Y}_B \lambda_B^*) \\ \nu^{*T}(-\bar{X}_0 + s_h^-) &= \nu^{*T}(-\bar{X}_B \lambda_B^*) \end{aligned} \quad (6)$$

and

$$\mu^{*T}(\bar{Y}_0 + s_h^+) - \nu^{*T}(-\bar{X}_0 + s_h^-) = (\mu^{*T} \bar{Y}_B - \nu^{*T} \bar{X}_B) \lambda_B^* \geq 0$$

as required to verify basis optimality in the new problem. Thus, the "projection" of the $(-\bar{X}_0, \bar{Y}_0)$ of DMU_0 onto $(-\bar{X}_0 + s_h^-, \bar{Y}_0 + s_h^+)$ is a projection onto the efficiency facet spanned by the efficient DMU's which determined DMU_0 's efficiency.

COMPUTATION OF EFFICIENCY

To compute the relative efficiency of the DMU's, we suggest using the DEA side of (4.2) for the n linear programs to be computed. Notice that they differ from one another only in the right hand side vector, a simplification from the computational situation for the CCR ratio model.

As before, however, we can immediately specify a feasible basis for DMU_0 . Here it differs from a diagonal matrix only in one column,

$$\begin{bmatrix} \bar{Y}_0 & -\bar{I} & 0 \\ -\bar{X}_0 & 0 & -I \end{bmatrix} \quad (8)$$

where $\bar{I}$ designates an $(s-1) \times (s-1)$ identity matrix with an additional first row of zeroes. We note, parenthetically, that the inverse of (8) can be written explicitly as

$$\begin{bmatrix} \bar{Y}_0 y_{10}^{-1} & \bar{I} & 0 \\ -\bar{X}_0 y_{10}^{-1} & 0 & -I \end{bmatrix} \quad (9)$$

where $\bar{Y}_0$ denotes Y_0 with y_{10} replaced by unity.

To speed up computation while making accessible the efficient facet information corresponding to DMU_0 , we point out that by Theorem 2 on reaching an optimal basis for DMU_0 , any DMU_k in this basis is efficient and its dual evaluators are those corresponding to DMU_0 ! Thus we do not need to separately compute an optimal tableau for DMU_k . We need only record the slacks for DMU_0 , its dual evaluator vector and what its subset of efficient DMU_k 's is and proceed to compute log-efficiency for the remaining unrecorded DMU's. Further, any DMU_p not in the basis, but which has zero " $(z_p - c_p)$ " is also in the same efficient facet. For, it is an alternate optimum which can be pivoted into the basis. But, since the old dual evaluators are still dual evaluators for it, no additional computation need be made. Thus, such DMU_p 's can also be recorded as corresponding to DMU_0 .

REFERENCES

1. A. Charnes, W. W. Cooper and E. Rhodes, Measuring the efficiency of decision making units. *European J. Op. Res.* 2, 429-444 (1978).
2. A. Charnes and W. W. Cooper, Management science relations for evaluation and management accountability. *J. Enterprise Management* 2, 143-162 (1980).
3. A. Charnes, W. W. Cooper and E. Rhodes, Evaluating program and managerial efficiency: an application of data envelopment analysis to program follow through. *Management Sci.* 27, 668-697 (1981).

CP 337
(CCS 441)

INVARIANT MULTIPLICATIVE EFFICIENCY AND PIECEWISE COBB-DOUGLAS ENVELOPMENTS *

A. CHARNES, W.W. COOPER and L. SEIFORD

Business - Economics Building, 203E, University of Texas at Austin, Austin, TX 78712, U.S.A.

J. STUTZ

University of Miami, Coral Gables, FL 33124, U.S.A.

Received November 1982

Revised April 1983

A new multiplicative efficiency formulation is developed wherein the efficiency values are invariant under changes in the units of measurement of outputs and inputs. It is shown that the associated Data Envelopment Analysis (DEA) implies that optimal envelopments are of piecewise Cobb-Douglas type. This leads to a new method for estimating frontier production functions of Cobb-Douglas type.

Multiplicative efficiency, log efficiency, data envelopment analysis, piecewise Cobb-Douglas forms, production functions

1. Introduction

In [1], Charnes, Cooper, Seiford and Stutz (C^2S^2) develop a multiplicative (or log) measure of the relative efficiency of multiple input, multiple output productive (or 'decisionmaking') units (DMU's). In contrast to the CCR measure [2,3], the multiplicative measure obtained in [1] is not invariant under change of units in the inputs or outputs. We show here how by a simple change preserving the multiplicative format a units invariant multiplicative measure can be obtained. Interestingly, the Data Envelopment Analysis (DEA) associated with this new modification necessarily yields optimal envelopments by Cobb-Douglas functions, i.e., the efficiency surface is piecewise Cobb-Douglas rather than merely log-linear! This uncovers a new role for Cobb-Doug-

las functions¹ - they are necessary for the units invariant property of a multiplicative measure.

2. Units invariant multiplicative efficiencies

The C^2S^2 multiplicative model reduces the input-output quantities to single virtual output to input ratios. If we now introduce an additional *virtual* output multiplier and *virtual* input multiplier, we obtain the following form for our problem to measure the efficiency of DMU, relative to all the n DMU's:

$$\begin{aligned} \max \quad & \left(e^{\hat{\eta}} \prod_{r=1}^s \gamma_{r0}^{\mu_r} \right) / \left(e^{\hat{\xi}} \prod_{i=1}^m X_{i0}^{\nu_i} \right), \\ \text{s.t.} \quad & \left(e^{\hat{\eta}} \prod_{r=1}^s \gamma_{rj}^{\mu_r} \right) / \left(e^{\hat{\xi}} \prod_{i=1}^m X_{ij}^{\nu_i} \right) \leq 1, \quad j = 1, \dots, n, \end{aligned} \quad (1)$$

with variables

$$-\hat{\eta} \leq 0, \quad -\hat{\xi} \leq 0, \quad -\mu_r \leq -\delta, \quad -\nu_i \leq -\delta \quad \forall r, i,$$

¹ Other properties relating Cobb-Douglas forms to more general classes of functions are examined in [5].

* This research is partly supported by ONR Contract N00014-81-C-0236, USARI Contract MDA-903-81-C-0365, AF Contract F49642-82-CO129 with the Center for Cybernetic Studies, The University of Texas at Austin; also ONR Contract N00014-80-0131 with Carnegie-Mellon University School of Urban and Public Affairs. Reproduction in whole or in part is permitted for any purpose of the United States Government.

X_{ij} the i th input of DMU $_j$, Y_{rj} the r th output of DMU $_j$, where $\delta > 0$ and DMU $_0$ is one of the n DMU's in the constraints.

The $\hat{\eta}$ and $\hat{\xi}$ are respectively the additional virtual output and virtual input 'multipliers' (exponents would be a more accurate designation) for the added output 'e' and input 'e', while the μ_r and ν_i are those for outputs Y_{rj} , $j = 1, \dots, n$, and inputs X_{ij} , $j = 1, \dots, n$, respectively. The outputs and inputs are all positive. The rationale of the 'exponentiation' and multiplication of these to obtain single 'virtual output' to 'virtual input' ratios as well as the relation of this multiplicative measure to the earlier 'additive' CCR measure is more fully explained in [1].

Suppose the units in the outputs and the inputs are changed so that Y_{rj} becomes $a_r Y_{rj}$ and X_{ij} becomes $b_i X_{ij}$ where $a_r, b_i > 0 \forall r, i$. (Note a_r or $b_i = 1$ corresponds to no change in those units.) Problem (1) becomes

$$\begin{aligned} \max \quad & e^{\hat{\eta}} \left(\prod_r a_r^{\mu_r} \right) \prod_r Y_{r0}^{\mu_r} / e^{\hat{\xi}} \left(\prod_i b_i^{\nu_i} \right) \prod_i X_{i0}^{\nu_i}, \\ \text{s.t.} \quad & e^{\hat{\eta}} \left(\prod_r a_r^{\mu_r} \right) \prod_r Y_{rj}^{\mu_r} / e^{\hat{\xi}} \left(\prod_i b_i^{\nu_i} \right) \prod_i X_{ij}^{\nu_i} \leq 1, \\ & j = 1, \dots, n, \\ & -\hat{\eta} \leq 0, -\hat{\xi} \leq 0, -\mu_r \leq -\delta, \\ & -\nu_i \leq -\delta \forall r, i. \end{aligned} \quad (2)$$

If (1) has optimal value $E(1)$ with $\eta^*, \xi^*, \mu^*, \nu^*$ an optimal solution, then we have $\exp(\hat{\eta}) = K \exp(\hat{\eta}^*) / \prod_r a_r^{\mu_r^*}$, $\exp(\hat{\xi}) = K \exp(\hat{\xi}^*) / \prod_i b_i^{\nu_i^*}$, μ^*, ν^* (where $K > 0$ assures $\hat{\eta}, \hat{\mu} > 0$) is 'feasible' for (2) with value $E(1)$. Hence for (2) the optimal value $E(2) \leq E(1)$. Similarly from an optimal solution $\hat{\eta}, \hat{\xi}, \hat{\mu}, \hat{\nu}$ to (2) we construct a feasible solution to (1) with value $E(2)$. Thereby $E(1) \leq E(2) \leq E(1)$, i.e. the efficiency value is invariant under change of units.

The Cobb-Douglas property

Taking logarithms in (1) and going to vector matrix notation as in [1], we obtain the dual linear programming problems:

$$\begin{aligned} \max \quad & \hat{\eta} - \hat{\xi} + \mu^T \hat{Y}_0 - \nu^T \hat{X}_0, \\ \text{s.t.} \quad & \hat{\eta} e^T - \hat{\xi} e^T + \mu^T \hat{Y} - \nu^T \hat{X} \leq 0, \end{aligned}$$

$$-\hat{\eta}, -\hat{\xi}, -\mu^T < 0; -\mu^T, -\nu^T \leq \delta e^T; \quad (3-I)$$

$$\begin{aligned} \min \quad & -\delta e^T s^+ - \delta e^T s^-, \\ \text{s.t.} \quad & e^T \lambda - \theta^+ = 1, -e^T \lambda - \theta^- = -1, \\ & \hat{Y} \lambda - s^+ = \hat{Y}_0, -\hat{X} \lambda - s^- = -\hat{X}_0, \\ & \lambda, \theta^+, \theta^-, s^+, s^- \geq 0. \end{aligned} \quad (3-II)$$

Here (3-II) represents the DEA side of the efficiency problem. Adding the first two equations in (3-II), we obtain $-\theta^+ - \theta^- = 0$. Since $\theta^+, \theta^- \geq 0$ we must have $\theta^+ = \theta^- = 0$. Thus (3-II) reduces to

$$\begin{aligned} \min \quad & -\delta e^T s^+ - \delta e^T s^-, \\ \text{s.t.} \quad & \hat{Y} \lambda - s^+ = \hat{Y}_0, -\hat{X} \lambda - s^- = -\hat{X}_0, e^T \lambda = 1, \\ & \lambda, s^+, s^- \geq 0. \end{aligned} \quad (4)$$

Thereby we have $\hat{Y}_0$ and $\hat{X}_0$ enveloped by convex combinations of the $\hat{Y}_j, \hat{X}_j$. With optimal solutions $\lambda^*, s^{*+}, s^{*-}$, we can write

$$Y_0 = \sum_{j=1}^n Y_j^{\lambda_j^*} e^{-s_j^{*+}}, \quad X_0 = \sum_{j=1}^n X_j^{\lambda_j^*} e^{-s_j^{*-}},$$

where $\sum_j \lambda_j^* = 1$, and by $Y_j^{\lambda_j^*}$, resp. $X_j^{\lambda_j^*}$, we mean $(Y_{1j}^{\lambda_j^*}, \dots, Y_{sj}^{\lambda_j^*})^T$ resp. $(X_{1j}^{\lambda_j^*}, \dots, X_{mj}^{\lambda_j^*})^T$. Thus our optimal envelopments are by Cobb-Douglas functions with $\lambda_j^* > 0$ implying that DMU $_j$ is efficient, i.e., DMU $_0$ is associated with the efficiency surface 'facet' spanned by those DMU $_j$'s for which $\lambda_j^* > 0$.

We note further that the simplified dual programs corresponding to (3) are now

$$\begin{aligned} \max \quad & \mu^T \hat{Y}_0 - \nu^T \hat{X}_0 + \omega, \\ \text{s.t.} \quad & \mu^T \hat{Y} - \nu^T \hat{X} + \omega e^T \leq 0, \\ & -\mu^T \leq -\delta e^T, -\nu^T \leq -\delta e^T; \\ \min \quad & -\delta e^T s^+ - \delta e^T s^-, \\ \text{s.t.} \quad & \hat{Y} \lambda - s^+ = \hat{Y}_0, -\hat{X} \lambda - s^- = -\hat{X}_0, e^T \lambda = 1, \\ & \lambda, s^+, s^- \geq 0. \end{aligned} \quad (6-I)$$

(6-II)

These results present us with a new method for estimating piecewise Cobb-Douglas production functions directly from empirical data. The form of (6-II) in contrast to that of [4] is also sufficiently simple that one can anticipate that the mathematical statistics of this type of Cobb-Douglas estimation may well be developed in the near future (see also the Appendix in [3]).

$$-v^T \leq \delta e^T; \quad (3-I)$$

$$\begin{aligned} \theta^- &= -1, \\ \theta^- &= -\hat{X}_0, \end{aligned} \quad (3-II)$$

DEA side of the first two equations. Since $\theta^+, \theta^- \geq 0$ (3-II) reduces to

$$- \hat{X}_0, e^T \lambda = 1, \quad (4)$$

developed by convex optimal solutions

$$X_j^*,$$

By X_j^* , we mean $(X_j^*)^T$. Thus our optimal Cobb-Douglas function MU_j is efficient, efficiency surface for which $\lambda_j^* > 0$. Simplified dual problem

$$(6-I)$$

$$e^T;$$

$$- \hat{X}_0, e^T \lambda = 1.$$

$$(6-II)$$

new method for as production data. The form is also sufficient that the use of Cobb-Douglas developed in the [3].

References

- [1] A. Charnes, W.W. Cooper, L. Seiford and J. Stutz, "A multiplicative model for efficiency analysis", *Socio-Economic Planning Sciences* (forthcoming 1983); also CCS 416, Center for Cybernetic Studies, University of Texas, Austin, TX (November 1981).
- [2] A. Charnes, W.W. Cooper and E. Rhodes, "Measuring the efficiency of decisionmaking units," *European J. Oper. Res.* 2, 429-444 (1978).
- [3] A. Charnes and W.W. Cooper, "Management science relations for evaluation and management accountability", *J. Enterpr. Management* 2, 143-162 (1980).
- [4] R.D. Banker, A. Charnes, W.W. Cooper and A.P. Schinnar, "A bi-extremal principle for frontier estimation and efficiency evaluations", *Management Sci.* 17 (12), 1370-1382 (1981).
- [5] A. Charnes, W.W. Cooper and A.P. Schinnar, "A theorem on homogeneous functions and extended Cobb-Douglas forms", *Proc. Nat. Acad. Sci. U.S.A.* 73 (10), 3747-3748 (1976).

APPENDIX C

1. "Axiomatics of Efficiency Reference Sets," by Charnes, Cooper, Rousseau and Semple.
2. Galley copy of article by Banker, Charnes, Cooper and Maindiratta.
3. Journal of Marketing article by Charnes, Cooper, Learner and Phillips.