Measuring the efficiency of decision making units

A. CHARNES
The University of Texas, Austin TX, U.S.A.

W.W. COOPER
Graduate School of Business Administration, Harvard University, Cambridge MA, U.S.A.

E. RHODES
State University of New York at Buffalo, Albany NY, U.S.A.

A nonlinear (nonconvex) programming model provides a new definition of efficiency for use in evaluating activities of not–for–profit entities participating in public programs. A scalar measure of the efficiency of each participating unit is thereby provided, along with methods for objectively determining weights by reference to the observational data for the multiple outputs and multiple inputs that characterize such programs. Equivalences are established to *ordinary* linear programming models for effecting computations. The duals to these linear programming models provide a new way for estimating extremal relations from observational data. Connections between engineering and economic approaches to efficiency are delineated along with new interpretations and ways of using them in evaluating and controlling managerial behavior in public programs.

This paper represents a revision of earlier versions entitled, respectively, 'Expositions, Interpretations, and Extensions of the Farrell Efficiency Measure', and 'Measuring the Efficiency of Decision Making Units With Some New Production Functions and Estimation Methods'. See [11], [12] and also [9]. This research was partly supported by NSF Grant No. SOC76-15876 'Collaborative Research on the Analytical Capabilities of a Goals Accounting System'. It is also part of the program of the Navy Personnel Research and Development Center and is partly supported by Project NR 047-021, ONR Contract N00014-75-C-0616 with the Center for Cybernetic Studies, The University of Texas, and ONR Contract N00014-76-C-0932 at Carnegie-Mellon University School of Urban and Public Affairs.

European Journal of Operational Research 2 (1978) 429–444.

1. Introduction

This paper is concerned with developing measures of 'decision making efficiency' with special reference to possible use in evaluating public programs. As we shall use it, the term 'program' will refer to a collection of decision making units (DMU's) with common inputs and outputs. These outputs and inputs will usually be multiple in character and may also assume a variety of forms which admit of only ordinal measurements. For example, in an educational program like 'Follow Through' [1], the efficiency of various schools, viewed as DMU's in this program, may be measured by reference to outputs involving the standard education categories: viz., cognitive, affective and psycho-motor skills via, respectively, (1) arithmetic scores, (2) psychological tests of student attitudes, e.g., toward the community and (3) student ability to understand and control bodily motions, e.g., by observing their ability to tread water and turn from front to back (and vice versa) in a swimming pool [2]. These are all to be regarded as 'valued' outputs even when there is no apparent market for them or even when other possible sources for reasonably supportable systems of weights are not readily available. The inputs may similarly range from fairly easy to measure (and weight) quantities like 'number of teacher hours' and extend to more difficult ones line 'time spent in program activities by community leaders and/or parents'.

Our use of terms like 'DMU' (decision making unit) and 'programs' will help to emphasize that our interest is centered on decision making by not-for-profit entities rather than the more customary 'firms' and 'industries'. It will also help us to emphasize that our data (as in the above example) are not readily weighted by reference to market prices [3] and/or other economic *desiderata* – such as costs of producing income earning capacity in students, with related rates of discount – in accordance with the

[1] A discussion of this Federally sponsored program which includes a use of the efficiency measures we shall be discussing may be found in [21]. This includes a use of various statistical tests of significance (using the so-called Kullback-Leibler statistic), which will not be discussed in the present paper.

[2] See the discussion in [3] for a use of measures like these in program- planning- budgeting (PPBS) contexts.

[3] We are referring to *actual* market prices and costs. Later we shall show how to obtain estimates of (optimal) production coefficients and relate them to theoretical (opportunity) costs and prices.

ways in which some public sector activities are sometimes evaluated.

Naturally we shall want to relate our ideas to developments in economics. This will be done by reference to production functions and related concepts such as 'cost duality', etc. Although adaptations of these concepts will be needed we shall also try to indicate what is involved at suitable points in this paper. (See below, Section 6, for instance.)

We shall also want to relate our ideas to other disciplines, like engineering, which are also concerned with efficiency measurement. This will be done not only in the interests of greater unity but also in the interest of distinguishing between efficiencies associated with an underlying production 'technology' and those due to managerial decision making when the former can be identified and separated from the latter by, e.g., engineering characterizations.

Of course, when this cannot be done (the usual case in empirical economics) [4] we will need to rest content with the somewhat less satisfactory concept of 'relative efficiency'. The latter will be determined by reference to suitably arranged 'rankings' of the observed results of decision making by various DMU's in the same program (e.g., the different schools in program Follow Through) while allowing for the fact that different amounts of inputs (sometimes legally stipulated) may be involved so that, e.g., some DMU's are more like members of one subset and less like members of other subsets, etc., in the 'amounts' of particular inputs and outputs utilized.

The meaning and significance to be accorded these characterizations will be clarified in the sections that follow. First we shall introduce our proposed measures and models. Then we shall provide characterizations which are wholly computational. Relations to selected lines of ongoing research will be delineated, followed by methods of estimation and interpretation in terms of simple numerical illustrations and analytical characterizations. A concluding section will then summarize what has been done and point up relevant shortcomings along with possible further lines of development.

[4] Such separation is even more difficult in public sector programs such as education, public safety, etc., where the meaning of a 'technology' is likely to be more ambiguous than in the case of manufacturing in the private sector, and even many service operations.

2. Model and definition

Our proposed measure of the efficiency of any DMU is obtained as the maximum of a ratio of weighted outputs to weighted inputs subject to the condition that the similar ratios for every DMU be less than or equal to unity. In more precise form,

$$\max h_0 = \frac{\sum_{r=1}^{s} u_r y_{r0}}{\sum_{i=1}^{m} v_i x_{i0}} \tag{1}$$

subject to:

$$\frac{\sum_{r=1}^{s} u_r y_{rj}}{\sum_{i=1}^{m} v_i x_{ij}} \leqslant 1\,; \qquad j = 1, \ldots, n,$$

$$u_r, v_i \geqslant 0\,; \quad r = 1, \ldots, s\,; \quad i = 1, \ldots, m.$$

Here the y_{rj}, x_{ij} (all positive) are the known outputs and inputs of the jth DMU and the $u_r, v_i \geqslant 0$ are the variable weights to be determined by the solution of this problem – e.g., by the data on *all* of the DMU's which are being used as a reference set. The efficiency of one member of this reference set of $j = 1, \ldots, n$ DMU's is to be rated relative to the others. It is therefore represented in the functional, for optimization – as well as in the constraints – and further distinguished by assigning it the subscript '0' in the functional (but preserving its original subscript in the constraints). The indicated maximization then accords this DMU the most favorable weighting that the constraints allow.

For the DMU's which concern us, these x_{ij} and y_{rj} values, which are constants, will usually be observations from past decisions on inputs and the outputs that resulted therefrom. We can, however, replace some or all of these observations by theoretically determined values if we wish (and are able) to conduct our efficiency evaluations in that manner.

Consider, for instance, the following definition (quoted from [14]) from the field of combustion engineering – viz., 'efficiency is the ratio of the actual amount of heat liberated in a given device to the maximum amount which could be liberated by the fuel [being used]'. In symbols,

$$E_r = y_r / y_R$$

where

y_R = Maximum heat that can be obtained from a given input of fuel,

y_r = Heat obtained by the input being rated from the same fuel input.

Although the definition of efficiency varies from one engineering field to another, the one above captures the essentials – viz., the rating is relative to some maximum possibility so that, always, $0 \leqslant E_r \leqslant 1$.

We can also obtain the above defined E_r from (1) as follows. For any given input amount x substitution in (1) gives

$$\max h_0 = \frac{uy_0}{vx_0}$$

$$\text{s.t.} \quad \frac{uy_R}{vx_R} \leqslant 1\,,$$

$$\frac{uy_r}{vx_r} \leqslant 1\,,$$

$$u, v \geqslant 0,$$

where $r = 0$ in the functional designates that the latter is being rated.

Let u^*, v^* represent an optimal pair of values. Since $y_R \geqslant y_r$ and $x_R = x_r = x$ this implies $u^* y_R = v^* x_R$ and using $x_0 = x$ we then have the functional equal to y_r/y_R as required.

In common with most engineering definitions we have here confined our development to ratios of single ourputs and/or inputs, or weighted sums thereof. The latter may be determined, again by engineering considerations (e.g., efficient fuel combinations), which are ordinarily not available for the economic applications we are considering. Provided we have the indicated observations on inputs and outputs for individual DMU's, however, we can at least achieve 'relative efficiency' ratings along the lines that we have been suggesting. This is the way the rest of the paper will be developed although, as already indicated, we can also insert engineering or other data for such ratings, if we wish, in various combinations.

Note that our weightings, as above, are objectively determined to obtain a (dimensionless) scalar measure of efficiency in any case [5]. I.e., the choice of weights is determined directly from observational data subject only to the constraints set forth in (1). Under these observations and constraints no other set of common weights will give a more favorable rating relative to the reference set. Hence if a (relative) efficiency rating of 100% is not attained under this set of weights then it will also not be attained from any other set.

[5] Scaling and invariance properties which are dealt with in [12] – see also [21] – will not be discussed in this paper.

3. Reduction to linear programming forms

The above model is an extended nonlinear programming formulation of an ordinary fractional programming problem. We have elsewhere (in [10] and [7]) supplied a complete theory in terms of which fractional programming problems may be replaced with linear programming equivalents. We therefore propose to use that theory here to make the above formulation computationally tractable for the large numbers $j(n)$ of observations as well as the smaller numbers of inputs $i(m)$ and outputs $r(s)$ which are likely to be of interest at least in economics applications.

We shall do this in a way that should provide further conceptual clarity (and flexibility) and also facilitate our making contact with related developments in economics. First consider the following model which is the reciprocal (inefficiency) measure version of (1):

$$\min f_0 = \frac{\sum_{i=1}^{m} v_i x_{i0}}{\sum_{r=1}^{s} u_r y_{r0}} \tag{2}$$

subject to:

$$\frac{\sum_{i=1}^{m} v_i x_{ij}}{\sum_{r=1}^{s} u_r y_{rj}} \geqslant 1\,; \quad j = 1, ..., n\,,$$

$$v_i, u_r \geqslant 0\,.$$

Now we propose to replace these nonconvex nonlinear formulations with an *ordinary* linear programming problem. We therefore first consider

$$\max z_0 \tag{3}$$

subject to:

$$-\sum_{j=1}^{n} y_{rj}\lambda_j + y_{r0} z_0 \leqslant 0\,; \qquad r = 1, ..., s\,,$$

$$\sum_{j=1}^{n} x_{ij}\lambda_j \leqslant x_{i0}\,; \qquad i = 1, ..., m\,,$$

$$\lambda_j \geqslant 0\,; \qquad j = 1, ..., n\,.$$

Because (3) is an ordinary linear programming problem it has a linear programming dual which we can write as follows:

$$\min g_0 = \sum_{i=1}^{m} \omega_i x_{i0} \tag{4}$$

subject to:

$$-\sum_{r=1}^{s} \mu_r y_{rj} + \sum_{i=1}^{m} \omega_i x_{ij} \geqslant 0\,,$$

$$\sum_{r=1}^{s} \mu_r y_{r0} = 1\,,$$

$$\mu_r, \omega_i \geqslant 0\,.$$

Because of the structure of (4) one can recognize that it is equivalent to an ordinary linear fractional programming problem. (See [10] and [7].) In fact, utilizing the theory of linear fractional programming with the transformation

$$\omega_i = t v_i\,; \quad i = 1, ..., m\,,$$

$$\mu_r = t u_r\,; \quad r = 1, ..., s\,,$$

$$t^{-1} = \sum_r u_r y_{r0}\,,$$

which, with $t > 0$, gives explicitly

$$\min f_0 = \frac{\sum_{i=1}^{s} v_i x_{i0}}{\sum_{r=1}^{s} u_r y_{r0}} \tag{6}$$

subject to:

$$\sum_{i=1}^{m} v_i x_{ij} - \sum_{r=1}^{s} u_r y_{rj} \geqslant 0\,; \qquad j = 1, ..., n\,,$$

$$v_i, u_r \geqslant 0\,,$$

as the linear fractional programming equivalent of (4). By very evident manipulations, however, we can see that (6) is the same as (2). Hence we can use (4) to solve (6) and therefore (2) and (1) as well. Q.E.D.

We are now in an advantageous position from several standpoints. We have a completely symmetric definition of efficiency which generalizes single output ratio definitions not only in economics but in engineering and other natural sciences. We do not need to solve the nonlinear (and nonconvex) problems in which these definitions are formalized. We need only solve the ordinary linear programming problem (4) in order to obtain both the optimal f_0^* or h_0^* and the weights $v_i^*, u_r^* \geqslant 0$, since the change in variables does not alter the value of the functional.

Thus,

$$f_0^* = g_0^* = z_0^* \tag{7.1}$$

and therefore

$$h_0^* = 1/z_0^*\,. \tag{7.2}$$

Also we have the wanted relative weights. Thus nothing more is required than the solution of (4) or (3) in order to determine whether $f_0^* > 1$ or, correspondingly, whether $h_0^* < 1$, with efficiency prevailing if and only if

$$f_0^* = h_0^* = 1\,. \tag{7.3}$$

We can also effect extensions in a variety of additional (new) directions [6]. Here, however, we prefer to make contact with various developments and also sketch a few of the ideas that are elsewhere described as Data Envelopment Analysis [7]. For this purpose we introduce

$$P_j = \begin{pmatrix} Y_j \\ X_j \end{pmatrix}; \qquad j = 1, ..., n\,, \tag{8}$$

wherein the subvector Y_j contains observed output values y_{rj}, $r = 1, ..., s$ for its components and the subvector X_j contains observed input values x_{ij}, $i = 1, ..., m$.

Now consider the following vector reformulation of (3):

$$\max z_0 \tag{9}$$

[6] For instance, we could utilize the duality theory that is now associated with fractional programming (as discussed in [18] and [25] – see also [5] and [7]) as distinguished from the duality theory of ordinary linear programming or what is sometimes called duality theory (see below) in cost and production theory.

[7] This is a method for adjusting data to prescribed theoretical requirements such as optimal production surfaces, etc., *prior* to undertaking various statistical tests for purposes of public policy analysis. See [21].

with

$$-\sum_{j=1}^{n} Y_j\lambda_j + Y_0 z_0 \leqslant 0\,,$$

$$\sum_{j=1}^{n} X_j\lambda_j \qquad \leqslant X_0\,,$$

$$\lambda_j \geqslant 0\,; \qquad j = 1, ..., n\,.$$

Let its optimal solution in the equivalent equation form with slack variables be represented by

$$z_0^*, s^{*+}, s^{*-}, \lambda_j^*\,; \qquad j = 1, ..., n\,, \tag{10}$$

where s^{*+} represents a vector of non-negative slack associated with the output inequalities and s^{*-} represents a vector of non-negative slack associated with the input inequalities. If $z_0^* > 1$ then via (7.1)–(7.3) the efficient frontier of the production possibility surface has *not* been attained.

Here, however, we can observe something more. If s^{*+} has any positive components then it is possible to increase the associated outputs in the amounts of these slack variables without altering any of the λ_j^* values and without violating any constraints. Similarly if s^{*-} has any positive components then we can reduce the inputs from X_0 to $X_0 - s^{*-}$ in an analogous manner. Thus, in either case the DMU being evaluated has not achieved (relative) efficiency even with $z_0^* = 1$. That is, unlike (1) and (2), the subsequent models for characterizing efficiency do not necessarily determine whether the DMU is efficient only by reference to the optimal functional value.

For ease of reference, we summarize what is involved for these latter cases as follows. No DMU can be rated as efficient unless the following conditions are both satisfied

(i) $z_0^* = 1$, and
(ii) The slack variables are all zero. (11)

It may be observed that these conditions are also the conditions for Pareto efficiency [8] – extended to cover production as well as consumption. Note that this assumes that a reduction in any input or an expansion in any output has some value. It does not require that these values be stipulated or prescribed in advance in any way. Indeed if efficiency measures are to be restricted to a scalar measure only, then objective computation of the weights from (4) [9], as already discussed, will suffice to produce what is wanted by direct substitution in (1).

[8] Also called Pareto–Koopmans efficiency. See Chapter IX in [8].

Now suppose we want to adjust all observations for purposes, say, of evaluating a *program*'s potential for a given DMU on the assumption that this program is efficiently managed by the specified DMU. This can be done by applying (11) in the following manner.

First, for a selected DMU we proceed via (9) to obtain the solution (10). Then we form a new problem from these data and their solution – viz.,

$$\max \hat{z}_0 \tag{12}$$

with

$$-\sum_{j=1}^{n} Y_j\hat{\lambda}_j + (Y_0 z_0^* + s^{*+})\,\hat{z}_0 \leqslant 0\,,$$

$$\sum_{j=1}^{n} X_j\hat{\lambda}_j \qquad \leqslant X_0 - s^{*-}\,,$$

$$\hat{\lambda}_j \geqslant 0\,; \qquad j = 1, ..., n\,.$$

We shall refer to (12) as the 'varied problem' and show that it may be used to eliminate all the inefficiencies detected in proceeding from (9) to (10). This includes (a) reducing inputs from the original X_0 vector of observations to the new (adjusted) input vector $X_0 - s^{*-}$, and also (b) increasing the originally observed output vector Y_0 to new (adjusted) values $(Y_0 z_0^* + s^{*+})$.

We now show that the thus adjusted observations satisfy the conditions for efficiency in (11) as follows. Evidently, we must have $\hat{z}_0^* \geqslant 1$ since $\hat{z}_0^* = 1$ in (12) together with (10) gives us the already secured optimal solution to (9). Now suppose we could have $\hat{z}_0^* > 1$ in (12). This would yield

$$-\sum_{j=1}^{n} Y_j\hat{\lambda}_j^* + Y_0\hat{z}_0^* z_0^*$$

$$\leqslant -\sum_{j=1} Y_j\hat{\lambda}_j^* + (Y_0 z_0^* + s^{*+})\,\hat{z}_0^* \leqslant 0\,,$$

$$\sum_{j=1}^{n} X_j\hat{\lambda}_j^* \leqslant X_0 - s^{*-} \leqslant X_0\,,$$

[9] When an extreme point method, such as the simplex method, is used then *either* (3) or (4) may be used since, as is well known, these methods simultaneously produce optimal solutions to *both* problems. See, e.g., [8].

since s^{*+} and s^{*-} are both non-negative. Evidently, the expressions on the left then satisfy the 'unvaried problem' (9) with $\hat{z}_0^*$ in place of z_0^*, and $\hat{\lambda}_j^*$ in place of λ_j^*. However, then also

$$\operatorname{Max} z_0 \geqslant z_0^* \hat{z}_0^* > z_0^*$$

when $\hat{z}_0^* > 1$. But $z_0^* = \max z_0$, by hypothesis. Thus a contradiction occurs which proves that $\hat{z}_0^* = 1$ is the optimal value for the varied problem (12).

Now we want to show that the optimal solution, $\lambda_j^*, j = 1, ..., n$, to the unvaried problem (9) is an optimal solution to the varied problem (12) with zero slack, i.e., the vectors $\hat{s}^{*+}$ and $\hat{s}^{*-}$ have zeros in all components as required for efficiency. First, via (10)

$$-\sum_{j=1}^{n} Y_j\lambda_j^* + Y_0 z_0^* + s^{*+} = 0$$

$$\sum_{j=1}^{n} X_j\lambda_j^* = X_0 - s^{*-} .$$

Thus λ_j^* is a feasible solution of the varied problem with $\hat{z}_0 = 1$. That is

$$-\sum_{j=1}^{n} Y_j\lambda_j^* + (Y_0 z_0^* + s^{*+})\hat{z}_0 = 0$$

$$\sum_{j=1}^{n} X_j\lambda_j^* = X_0 - s^{*-}$$

with $\hat{z}_0 = 1$. It is also optimal since as we have just shown, $\hat{z}_0^* = 1$. Further, the optimal slacks $\hat{s}^{*+}$ and $\hat{s}^{*-}$ are all zero. Q.E.D.

In short, the indicated adjustments do, in fact, always bring the original observations into the relevant efficient production set. No new computations are required after the z_0^*, s^{*-} adjustments are effected for the original Y_0, X_0 data for the efficiency comparisons we may subsequently want to make.

As we shall shortly see – in section (6) below – we can use these results to obtain a surface corresponding to a well-defined relation between output and inputs. For the single output case this relation corresponds to a function in which output is maximal for all the indicated inputs. It therefore formally fulfills the requirements of a 'production function' or, more generally, a 'production possibility surface' [10] in the case of multiple outputs. In this manner we obtain a new type of production function which has a variety of advantages that we shall indicate as we proceed. Here we may adumbrate some of these advantages in a summary way as follows. Unlike other types of production functions, this one derives from (and is therefore directly applicable to) empirical observations. It also bypasses the intractable problems of aggregations associated with other types of production functions [11] and, finally, it lends itself to comparative statics for such purposes as determining whether technological change is occurring. These 'comparative' static uses may be accomplished in various ways such as adopting the convention that the same DMU is to be regarded as a different entity in each relevant time period.

By means of these production function concepts and the procedures we shall associate with them, the requirements of economic theory may then be brought to bear in new and modified ways for public policy evaluations, and, at the same time, provide a variety of new predictions and control possibilities for *program* managers – i.e., managers, legislators, etc., who have total program responsibility. For example, it then becomes possible to distinguish between the 'program efficiency' – that may be predicted with efficient management – and to distinguish this from other predictions (and evaluations) that might be effected on the assumption that all managers will continue to operate only at past levels of efficiency [12]. Similarly, it is possible to allow for predictions of future changes in technology instead of assuming that a static longrun production function has been achieved – as is presently being done in many of the extant studies for energy policy guidance in Western countries.

This may all be accomplished, we should note, without interfering with subsequent statistical testing and evaluation. As in the DEA approach which we discuss elsewhere, these statistical tests, which are

[10] We are using this term in these sense of the activity vectors discussed in Arrow and Hahn [4]. Indeed, in Section 5, below, we shall explicitly show how our duality characterization may be used to secure numerical estimates of these coefficient values (or, rather, their production function counterparts) from observational data.

[11] Disaggregation may be necessary, however, when data on individual DMU's are not available. See the guidelines supplied in [12] and [21]. For a discussion and an attempt to deal with the difficulties of aggregation in *other* types of production functions, see [24].

[12] See [21] for further discussion and a detailed application that distinguishes between 'program efficiency' and 'managenal efficiency' in 'Program Follow Through' of the U.S. Office of Education.

applied *after* the indicated adjustments, are often greatly simplified relative to other alternatives. Proceeding without first effecting such adjustments in the data would not only fail to utilize the underlying theory, it would also contradict the requirements of that theory unless (a) one can assume that *all* DMU's are operating efficiently or (b) one can supply some alternate method of allowing for observations which are *not* on the efficient production possibility frontier [13].

4. Isoquant analysis and Farrell efficiency

We proceed to still further implications of our definitions of efficiency and its production – economics – management consequences. For this we turn to the more familiar form of isoquant analyses, and related production function concepts. This will also allow us to make contact with the important work initiated by M.J. Farrell [15] [14].

First we will undertake the wanted clarification (and provide contact) via an isoquant analysis which corresponds to the one that Farrell used. Then we shall supply a model for generating the related efficient surface for the production function. We will also relate the latter to an associated cost function, and an extension of Shephard's lemma. Then we shall exhibit further relations between this and Farrell efficiency. In between we shall show how the duals (See [8], [12] and also [13].) to our models may be used to provide a new way of estimating production function coefficients either in their own right or in association with statistical estimation techniques (e.g., via the DEA approach outlined at the close of the last section).

To initiate these developments we first observe that we are now concerned with the case of only one output, which is the same for every DMU. In this case (3) specializes to

$$\max z_0 \tag{13}$$

subject to:

$$-\sum_{j=1}^{n} y_j\lambda_j + y_0 z_0 \leqslant 0\,,$$

$$\sum_{j=1}^{n} x_{ij}\lambda_j \leqslant x_{i0}\,; \qquad i = 1, ..., m$$

$$\lambda_j \geqslant 0\,; \qquad j = 1, ..., n$$

where, because we are concerned with only one output, the extra subscript is dropped on the y_j for each of the $j = 1, ..., n$ DMU's. Making the scaling change of variables $\lambda_j = \lambda_j'/y_j$, $z_0 = z_0'/y_0$, then dropping the primes and the constant multiplier $1/y_0$ in the functional, we obtain the equivalent form,

$$\max z_0 = \sum_{j=1}^{n} \lambda_j \tag{14}$$

subject to:

$$\sum_{j=1}^{n} x_{ij}'\lambda_j \leqslant x_{i0}'\,; \qquad i = 1, ..., m\,,$$

$$\lambda_j \geqslant 0\,; \qquad j = 1, ..., n\,,$$

wherein the values $x_{ij}' = x_{ij}/y_j$; $j = 1, ..., n$ and $x_{i0}' = x_{i0}/y_0$, are obtained from the change of variable operation we have just described.

We have made the transformation from (13) to (14) in the indicated manner to facilitate contact with the work of Farrell [15]. We shall also represent (14) equivalently as

$$\max z_0 = \sum_{j=1}^{n} \lambda_j \tag{14.1}$$

with

$$\sum_{j} P_j\lambda_j \leqslant P_0\,,$$

$$\lambda_j \geqslant 0\,; \qquad j = 1, ..., n$$

or its equivalent

$$\max z_0 = \sum_{j=1}^{n} \lambda_j \tag{14.2}$$

with

$$\sum_{j=1}^{n} P_j\lambda_j + \sum_{i=1}^{m} e_i s_i = P_0 \text{ and}$$

[13] For a discussion of problems involved in dealing with such estimation problems even in the case of a single output and very few inputs see [1] and [2].

[14] Examples involving continuation of this work may be found in [2] and [16].

[15] To make full contact with Farrell's work it is also necessary to deal with (and eliminate) some of his awkward concepts such as 'points at infinity', etc. This is done in [12] but in the interest of brevity we shall not attempt to repeat these developments here.

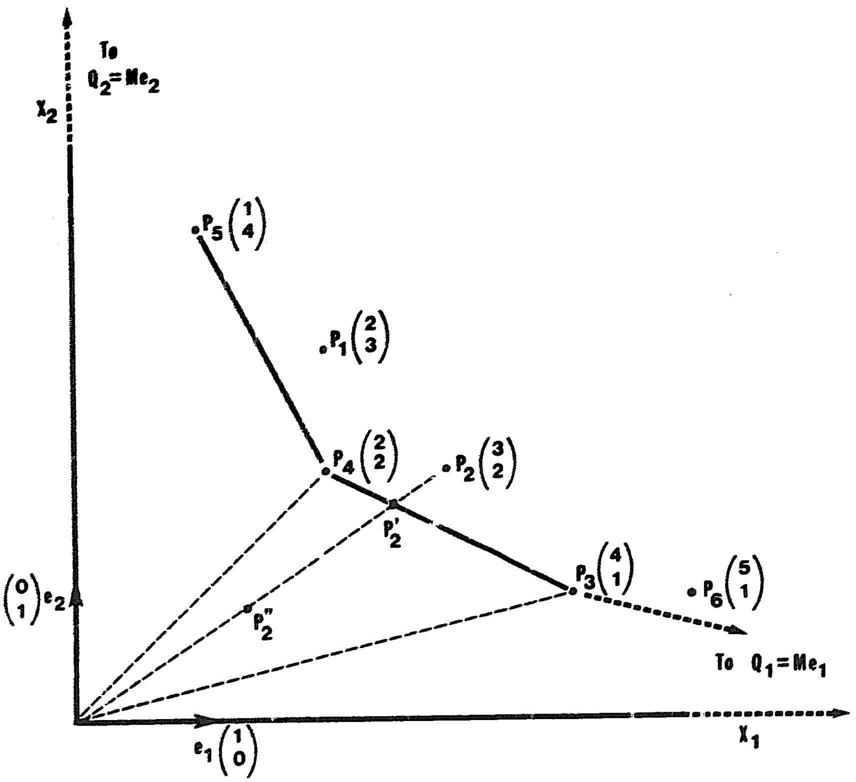

Fig. 1. Efficiency points and isoquant.

$\lambda_j, s_i \geqslant 0$,

wherein the P_j represent

(a) a normed version of (8), in which the components of X_j are divided by their corresponding output amounts, and

(b) the output is then eliminated from the vector and (in normed version) associated instead with the unit coefficients of the variables λ_j in the functional. In short, $P_j = X_j'$ with only normed inputs x_{ij}' for its components. The e_i are unit vectors with unity in row $i = 1, ..., m$ and zeros elsewhere and the s_i are 'slack' variables.

The conditions (11) continue to apply, of course, but now the referent for efficiency is only to inputs [16]. To see what this means refer to Fig. 1 on which is depicted a situation for six DMU's using two inputs normed on their respective output values. Here the DMU's are all portrayed for the case of two inputs normed on their respective outputs. E.g., P_5 has $x_{15}' = 1, x_{25}' = 4$ while P_3 has $x_{13}' = 3, x_{23}' = 2$, and so on.

Suppose it is now decided to test the efficiency of P_2. This can be done by inserting $P_2 \equiv P_0$ in (14.2) and utilizing any adjacent extreme point method, such as the simplex or dual methods [17], to obtain:

$$\tfrac{5}{6} P_4 + \tfrac{1}{3} P_3 = P_0 = P_2 \,.$$

I.e., these methods express P_2 in terms of an optimal basis with $\lambda_4^* = \frac{5}{6}, \lambda_3^* = \frac{1}{3}$ and all other $\lambda_j^* = 0$. Thus we have $z_0^*(P_2) = \frac{7}{6}$ so that, by reference to (11), P_2 is not efficient.

We can now give this measure of efficiency operational meaning by convexifying the above expresssion, which we do by dividing by $\frac{7}{6}$, to obtain:

$$\tfrac{5}{7} P_4 + \tfrac{2}{7} P_3 = \tfrac{6}{7} P_2 = P_2' \,.$$

As in Fig. 1 this expresses P_2' as a convex combination of the optimal basis vectors, P_4 and P_3, and hence brings the resulting P_2' point onto the line segment connecting them. In efficiency terms this means that if P_2 had been producing as efficienctly as P_2' – or, equivalently, as efficiently as the indicated convex combination of P_4 and P_3 – then it should have been

[16] We refer only to inputs in order to shorten the discussion that follows, although, of course, the manipulations can also be extended to possible output increases.

[17] See the discussion of these methods and an analysis of their special properties in Chapter XIII of [8].

able to secure its one unit of output with only $\frac{6}{7}$ of the amount of each of x_1 and x_2 that it was observed to utilize.

The line connecting P_4 and P_3 represents one segment of what Farrell refers to as the 'unit isoquant'. It is also efficient. That is, there cannot be a point such as P_2'' between this unit isoquant and the origin since that would imply that P_4 and P_3 were not an optimal basis [18].

The above analysis has employed two assumptions which we shall refer to as the 'isoquant' and 'ray assumptions', respectively. The latter, i.e., the ray assumption, which corresponds to assuming constant returns to scale [19], may be relaxed but at a cost in complications and resulting explanations which we shall not undertake here. The former, i.e., the isoquant assumption representation, is critical and it may not be relaxed in the sense that it provides the comparison point between P_2 and P_2' [20].

We have just described the operational significance that we wish to accord to our efficiency measure for the case of P_2. Now consider P_1 for which we obtain

$$\tfrac{5}{6}P_4 + \tfrac{2}{6}P_5 = P_1$$

when employing adjacent extreme point methods. That is, the optimal basis consists of P_4 and P_3 (which are both thereby characterized as efficient) with $z_0^*(P_1) = \lambda_4^* + \lambda_5^* = \frac{7}{6}$ so that, by virtue of (11), P_1 is also *not* efficient.

As can be seen, we have $z_0^*(P_1) = z_0^*(P_2) = \frac{7}{6}$. However, P_1 and P_2 are expressed in terms of different bases and hence have different referents. The convexification process used for P_2 is also employable for P_1 to obtain

$$\tfrac{5}{7}P_4 + \tfrac{2}{7}P_5 = \tfrac{6}{7}P_1 = P_1'$$

so that, also, the same ratio of contraction for all resources is needed to bring P_1 onto the efficient surface. However, it lies in the cone through the origin formed from P_4 and P_5 whereas P_2 is in the cone formed from P_4 and P_3. This condition (which arises from the non-negativity imposed on the admissible λ_j values) has an advantage for the sorts of public program applications we are considering. As observed in the opening section, many DMU's such as different school districts, etc., work under varying constraints with respect to inputs (as well as outputs) for the same program conducted in different locales or different parts of the country. Hence it is well to have the referents used for scoring the efficiency of each DMU as alike to it as possible, at least in some loose sense, while not interfering 'too much' with the wanted efficiency ratings [21].

We now conclude this section with the case of P_6 in Fig. 1, for which alternate optima are present since

$$1P_6 = P_6 \quad \text{and} \quad 1P_3 + 1\,e_1 = P_6$$

with $z_0^*(P_6) = 1$ in either case. Note, however, that the second of these two solutions has $s_1^* = 1$, which is to say that efficiency is not attained for P_6 until slack in the amount of $s_1^* = 1$ is subtracted from the first component in P_6 – after which it is coincident with P_3 in Fig. 1; see (11).

To avoid ambiguity, and to retain the wanted operational meaning for efficiency, it is necessary to maximize the slack values, but in a way that accords this a lower priority than maximizing the λ_j values. This is done by replacing (14.2) with

$$\max z_0 = \sum_{j=1}^{n} \lambda_j + \frac{1}{M}\sum_{i=1}^{m} s_i \tag{14.3}$$

subject to:

$$P_0 = \sum_{j=1}^{n} P_j\lambda_j + \sum_{i=1}^{m} e_i s_i\,,$$

$$\lambda_j, s_i \geqslant 0\,; \qquad j = 1, ..., n\,, \qquad i = 1, ..., m$$

where M is the usual large (non-Archimedean) 'quantity' (cf., e.g., [22]) which insures that $1/M > 0$ is always smaller than any positive real value that may be assumed by any λ_j.

[18] Proofs of propositions like these, which are fairly transparent, will not be given in the present paper. They may be found in [12]. The latter source may also be consulted for ways in which the formulations being used here differ from the one suggested to Farrell by A. Hoffman in the discussion associated with [15] and subsequently employed by Farrell and Fieldhouse in [16].

[19] Note, however, that this constant will be different, in general, for every P_j.

[20] It is to be understood, however, that our analyses are also applicable when various transformations are utilized to bring other functions such as, e.g., Cobb-Douglas functions, into suitable piecewise linear representations.

[21] Note that both P_1 and P_2 are also in the cone formed from P_5 and P_3 in Fig. 1 but the line segment connecting P_3 and P_5 is *not* efficient – and hence P_3 and P_5 will not be an optimal basis unless one wants to impose further constraints on the basis choices. See the discussion in Section 2, above.

5. Duality relations for coefficient estimation

Farrell distinguished between the above efficiency[22] which he referred to as 'technical efficiency' and other types of efficiency which he referred to as 'price efficiency' and 'overall efficiency' – with the latter being characterized as involving both 'price *and* technical efficiency'. Here we may note that Farrell restricted his studies mainly to technical efficiency – and for the reasons set forth in our introduction we shall do the same. Concerning price efficiency (in the sense of 'actual market prices') Farrell contented himself, for the most part, with pointing to the formidable difficulties involved in assessing even *relative* price efficiency, e.g., because of the varying motives of buyers and sellers. One may, however, come at this problem from the standpoint of 'in principle prices' and/or 'opportunity costs' such as are obtainable via standard economic theorems. This, in any case, is the route we shall follow as we also show how the dual to the above linear programming formulations may be used to secure the values of the slope coefficients for the efficient isoquants.

Of course, we shall generally be dealing with m-dimensional representations in which line segments such as those portrayed in Fig. 1 are replaced by efficient 'facets'. We have, however, now provided a model and a method [23] for generating these factors for any finite number of inputs since these facets correspond to all of the convex combinations of points that can be generated from the optimal bases. Specification of these optimal bases thus constitutes one way of representing these facets.

Another way is available by reference to the duality relations of linear programming as we shall now show. We therefore write the dual to (14.1) as

$$\min g_0 = \omega^{\mathrm{T}} P_0 \tag{15}$$

with

$$\omega^{\mathrm{T}} P_j \geq 1 \; ; \quad j = 1, ..., n ,$$

$$\omega^{\mathrm{T}} \geq 0 ,$$

where the superscript T represents transposition, as usual, so that, e.g., ω^{T} represents the transpose of the column vector ω with components $\omega_1, ..., \omega_m$ and $\omega^{*\mathrm{T}}$ denotes an optimum vector for these variables in the above problem.

We now observe that $\omega^{*\mathrm{T}} P_i = 1$ for each P_i in an optimal basis. To obtain our alternate representation of this facet in terms of the slopes of the efficient isoquant surface we therefore need only show that $\omega^{*\mathrm{T}}$ is orthogonal (see [20] and/or [8]) to the efficient facet spanned by these P_i. To do this it suffices to show that ω^* is orthogonal to any direction lying in the facet, e.g., to any vector which is the difference, $\bar{P} - \bar{\bar{P}}$, of two vectors in the facet. Since, by assumption, $\bar{P}$ and $\bar{\bar{P}}$ are in the facet, we can express them in terms of these same P_i via

$$\bar{P} = \sum_i P_i \bar{v}_i , \qquad \bar{\bar{P}} = \sum_i P_i \bar{\bar{v}}_i , \tag{16.1}$$

$$0 \leq \bar{v}_i , \bar{\bar{v}}_i \leq 1 , \qquad \sum_i \bar{v}_i = \sum_i \bar{\bar{v}}_i = 1$$

where summation is over the indexes of these P_i. But then

$$\begin{aligned} \omega^{*\mathrm{T}}(\bar{P} - \bar{\bar{P}}) &= \sum_i (\omega^{*\mathrm{T}} P_i \bar{v}_i - \omega^{*\mathrm{T}} P_i \bar{\bar{v}}_i) \\ &= \sum_i (1 \, \bar{v}_i - 1 \, \bar{\bar{v}}_i) \\ &= \sum_i \bar{\bar{v}}_i - \sum_i \bar{\bar{v}}_i = 0 , \end{aligned} \tag{16.2}$$

since $\Sigma_i \bar{v}_i = \Sigma_i \bar{v}_i = 1$. Q.E.D. Hence, the ω^* corresponding to this efficient facet determined by this optimal basis is orthogonal (or normal) to it. Thus, ω^* is normal to the hyperplane containing this facet. The equation of this hyperplane is

$$\omega^{*\mathrm{T}} x = 1 \; ; \tag{17}$$

i.e., any x satisfying this equation is a point in the linear space spanned by the totality of the P_j's and e_i's. The portion of the hyperplane consisting of the

[22] The above representation, given at the conclusion of the preceding section, implicitly retains concepts like 'points at infinity', – e.g., as represented by Q_1 and Q_2 in Fig. 1 – but we lay aside the development necessary to eliminate them. In any case, we may have recourse to (1) for the wanted scalar measure of efficiency without reference to these considerations. I.e., $h_0^* = 1$ then suffices to determine whether or not the corresponding P_0 is efficient.

[23] Any adjacent extreme point method will do.

efficient facet is the set of all x which are convex combinations of the P_i which form the optimal basis.

Of course, we will need to obtain the generality different ω^{*T} for the 'slopes' associated with the different facets of the efficient surfaces and the ranges for which they are valid. These ω^{*T} values, however, are obtainable within the computational process for the λ^*. For, as is well known (see, e.g. [8]), adjacent extreme point methods (such as the simplex and dual methods) simultaneously solve both (14.1) and (15). That is, the optimal λ^* and ω^{*T} values are obtained from the *same* tableau. Hence the solution to one of these problems also provides the solution to the other without extra computational effort.

For instance, the tableau that provided the solution and hence the optimal basis for characterizing the efficiency of P_1 with $\lambda_4^* = \frac{5}{6}, \lambda_5^* = \frac{2}{6}$ – see preceding section – also provides the associated dual variables $\omega_1^* = \frac{1}{3}, \omega_2^* = \frac{1}{6}$. But, as previously noted, we can express any $x^T = (x_1, x_2)$ in this facet via the basis vectors, P_4, P_5, as

$$x = P_4\nu_4 + P_5\nu_5$$

with $\nu_4, \nu_5 \geqslant 0$ and $\nu_4 + \nu_5 = 1$. Hence also

$$\omega^{*T}x = \omega_1^*x_1 + \omega_2^*x_2 .$$

For this segment we have $\omega_1^* = \frac{1}{3}, \omega_2^* = \frac{1}{6}$ and therefore

$$S[P_5, P_4] \equiv$$
$$\{(x_1, x_2): \tfrac{1}{3}x_1 + \tfrac{1}{6}x_2 = 1 ; 1 \leqslant x_1 \leqslant 2 , 2 \leqslant x_2 \leqslant 4\} ,$$

where the square brackets mean that the points P_5, P_4 designate this facet (=segment) with the relation prescribed for the ranges indicated in the curly brackets.

Similarly for $S[P_4, P_3]$ we have

$$S[P_4, P_3] \equiv$$
$$\{(x_1, x_2): \tfrac{1}{6}x_1 + \tfrac{1}{3}x_2 = 1 ; 2 \leqslant x_1 \leqslant 4 , 1 \leqslant x_2 \leqslant 2\} .$$

while the isoquant segment from P_3 to Q_1 (see Fig. 1) may be represented by

$$S[P_3, Q_1) \equiv$$
$$\{(x_1, x_2):\ 1 = \frac{1}{M}x_1 + x_2;\ 4 \leqslant x_1, 1 \leqslant x_2\}$$

where the square bracket indicates that P_3 is included while the parenthesis on the right indicates that $Q_1 = Me_1$, is excluded from the set and a similar characterization applies for points like Q_2, etc.

We now replace (17) with

$$S[P_i\text{: all } i \in I] \equiv \tag{17.1}$$
$$\{x:\ \omega^{*T} = 1;\ x = \sum_{i \in I} P_i\nu_i \text{ for all } \nu_i \geqslant 0, \sum_{i \in I} \nu_i = 1\}$$

in order to extend the preceding analysis to higher dimensions. Here the efficient surfaces are composed of facets and ω^{*T} is the normal to such a facet (a simplex) (see [8], p.242) which is the convex set spanned by the optimal basis. The theory that has been developed around the usual adjacent extreme point methods, including their nonlinear extensions allows us to obtain all of the needed information for interpretation as well as continuation from one efficient facet to another via the computational strategies that these methods allow in conjunction with each other [24].

To conclude this section we can obtain a straightforward way of interpreting the components of the normal vector, ω^{*T}, as marginal productivities in the indicated inputs. We do this by returning to the two dimensional case, such as the one depicted in Fig. 1. The condition for remaining on the same isoquant when effecting substitutions is

$$\omega_1^*x_1 + \omega_2^*x_2 = \hat{\omega}_1^*\hat{x}_1 + \hat{\omega}_2^*\hat{x}_2 ,$$

where $x \neq \hat{x}$ are two points on the indicated isoquant, but on possibly different segments with slope vectors ω^* and $\hat{\omega}^*$. Adding $\omega_1^*\hat{x}_1 - \omega_1^*\hat{x}_1 = 0$ on the left and $\omega_2^*\hat{x}_2 - \omega_2^*\hat{x}_2 = 0$ on the right, we then effect a series of algebraic manipulations to obtain [25]

$$\frac{\Delta x_2}{\Delta x_1} = -\frac{\omega_1^*}{\omega_2^*} - \frac{\Delta\omega_1}{\Delta x_1}\frac{\hat{x}_1}{\omega_2^*} - \frac{\Delta\omega_2}{\Delta x_1}\frac{\hat{x}_2}{\omega_2^*}$$

where

$$\Delta x_1 = x_1 - \hat{x}_1 ; \qquad \Delta x_2 = x_2 - \hat{x}_2$$
$$\Delta\omega_1 = \omega_1^* - \hat{\omega}_1^* ; \qquad \Delta\omega_2 = \omega_2^* - \hat{\omega}_2^* .$$

When both x and $\hat{x}$ are on the same isoquant segment we will have $\Delta\omega_1 = \Delta\omega_2 = 0$. Indeed, in this case we

[24] E.g., the simplex method of G.B. Dantzig and the dual method of C.E. Lemke may be joined together for these purposes. See [12] and [21] for further discussion along with numerical illustrations. For a discussion of the wealth of information relevant to economic and managerial (policy) interpretations available in the resulting tableaus, see Ch. VI in [8].

[25] This can be restated in terms of elasticities of substitutions, if desired.

will have a well defined derivative and so we may replace the above finite difference formulation with

$$\frac{dx_2}{dx_1} = -\frac{\omega_1^*}{\omega_2^*},$$

which is the usual expression relating the marginal rate of substitution between two inputs, as characterized by the derivative on the left, and the negative reciprocal of the ratio of their marginal productivities, on the right, with equality between the two sides required to stay on the same isoquant. In short, ω_i^* is the marginal productivity of the ith factor, $i = 1, 2$. Since, by the assumptions of economics, these marginal productivities are never negative, it follows that the marginal rate of substitutions is non-positive and the isoquant is also assumed to be convex (and continuous), etc., if we want to maintain contact with the usual theorems in these parts of economic theory.

6. Production functions and cost–price relations

We are now ready to construct the production function and cost relations from the just derived ω^* values. Before providing the model for doing this, however, we might sharpen some of the points that have already made by pausing to see what we have achieved and how it relates to other kinds of models and approaches in economics.

Our function is evidently derived from empirical observations. Although these observations are all at the level of the individual 'firms' [26], we evidently have something that differs from other studies and production function estimates at the individual firm level (see [19]) since, by hypothesis, we are considering *all* pertinent DMU's.

Approaches which include all such data have customarily been undertaken only at *aggregate* levels, with all the attendant difficulties (and assumptions) required to ensure that the functions thus estimated have the extremal properties that a 'production function' must possess [27]. Our functions evidently differ from the latter in that (a) the data are not aggregated prior to estimation and (b) the resulting estimates are optimal – i.e., they have the requisite extremal properties – to the extent that the data allow [28]. It follows that we are entitled to use these optimality properties in deducing the further theorems and relations that we shall note below [29] since they also refer to these same data or else to estimates derived therefrom via rigorously established optimizing methods and principles.

In some ways these production functions are reminiscent of ones that might be associated with Alfred Marshall's concept of a 'representative firm'. (Discussion in [28]; also [9].) Here, however, the referent is rather to 'representative efficient firms'. Note that the plural is required insofar as there is more than a single efficient facet. The continuum within each facet is then 'representative' of the efficiency for which the originally observed efficient firms serve as referents.

We commented earlier on the appropriateness of these 'representative' facets (and the cones generated from them) [30] for evaluating the efficiency of DMU's in various public programs. They provide an equally appropriate referent for making estimates of what various DMU's *should* be able to produce in the way of outputs [31] given the factor amounts and/or the relations between various inputs that may be prescribed for them. This is to say that we shall want to refer each DMU to its 'representative' facet (or cone) en route to making the translations and transformations prescribed in (12) for bringing them onto the relevant production function surface. Given the latter projections we are then in a position to estimate what outputs we may expect with efficient production from various resource allocations to each DMU in the programs being considered. These results may then be aggregated in a variety of ways for assessing or controlling the activities to be generated by these DMU's [32].

[26] We are using the term 'firms' interchangeably with DMU's in this section in order to avoid circumlocutions in relating these developments to pertinent parts of traditional economics. See Johnston [19].

[27] For a discussion of the onerous (and unrealistic) conditions needed to obtain these properties at aggregate levels, see Sato [24], pp. 3–8 ff.

[28] Including ex cathedra allowances, when available, as noted in Section 2.

[29] For a further critique of the use of these optimality properties in association with such aggregate functions (e.g., in some of the current energy policy studies) along with simple counteramples to the assumptions used in such studies see [13].

[30] See the discussion of 'representative convex sets' and the cones they 'represent' on p. 236 in [8].

[31] Including multiple outputs as discussed in Section 2, above, since the principles we are discussing extend in relatively obvious ways to the more general case of production possibility surfaces.

[32] The principles to be used in effecting these aggregations are set forth in [12].

To allow for all possible variations in such inputs and outputs we, of course, need to have a suitable way of obtaining the production function surface. We therefore now provide a way of doing this that relates our results to another type of 'duality' in economics [33].

The latter, as introduced by Samuelson [23] and Shephard [26,27], proceeds by introducing a cost function $C(y, p)$ as determined via

$$C(y, p) = \min p^{\mathrm{T}} x \text{ for } x \in L(y) \tag{18}$$

where

$$L(y) \equiv \{x\text{: at least the output vector } y \text{ is produced}\}.$$

In other words $L(y)$ is the point-to-set mapping $y \to L(y)$. For instance, in the single output case, as in Fig. 1, if it is obtained via the isoquant associated with $y = 1$ we would have $L(y)$ designating the set of all of the points $x = (x_1, x_2)$ interpreted as input combinations on or to the northeast of this isoquant. From a knowledge of these relations, the cost function $C(y, p)$ is then to be obtained via the indicated minimization where p is a 'price' vector with component p_i representing the 'price' per unit x_i, the amount of the ith factor input.

In our case we want our production function to be empirically based. That is, we want our production function to be based on observed input output data or estimates derived from them such that no firm from the observation set has a larger output for any inputs that may be specified. Also no non-negative combination of these firms can have a larger output when extrapolations or interpolations from the original observations must be undertaken for the specified input values.

To these ends we now proceed to obtain the wanted production function as follows. Let y be some prescribed output [34] and let a_s^{T}, the sth row of the matrix A, represent the set of coefficients associated with the sth efficient facet estimated from the data, as described in the preceding section. I.e., $a_{si} \equiv \omega_{si}^*$. See (17) or (17.1). Then also let P be a matrix with its column vectors P_j representing the observational data for each of the original $j = 1, ..., n$ DMU's [35].

[33] We have elsewhere suggested that these relations might better be associated with the mathematics of 'transform theory'. See [13].

[34] We can, of course, extend this to the multiple output case very readily, as in (18).

[35] One can omit the P_j which are not efficient or else one can adjust and bring them into the efficient set in the manner indicated in (12).

Then we write

$$\min p^{\mathrm{T}} x \tag{18.1}$$

with

$$Ax \geqslant ye\,,$$

$$-P\lambda + Ix = 0\,, \qquad \lambda \geqslant 0\,,$$

in which e is a column vector with all elements equal to unity while I is the identity matrix so that (i) $Ix = P\lambda$, $\lambda \geqslant 0$ assures us that we will be deriving our production function from empirical observations, and (ii) the way the vectors a_s were derived in $a_s^{\mathrm{T}} x \geqslant y$, together with the minimizing objective, assures us that we will always be on an efficient frontier when an optimum solution is obtained to (18.1) for any prescribed y.

The x^*, y values resulting from the solutions to (18.1) for different p are all points on the production function surface. When this surface is available one can evidently also proceed in the reverse manner, if desired, to estimate the output value that can be secured from any prescribed x.

Also, more is available since the above model can be used in various ways. Here, however, we mainly want to relate it to the cost-production function duality relations of economic theory. Therefore we write the mathematical programming dual to (18.1) as

$$\max y\eta^{\mathrm{T}} e \tag{18.2}$$

with

$$\eta^{\mathrm{T}} A + u^{\mathrm{T}} I = p^{\mathrm{T}}\,,$$

$$-u^{\mathrm{T}} P \leqslant 0\,, \qquad \eta^{\mathrm{T}} \geqslant 0\,.$$

Via the duality theorem of mathematical programming, we then have

$$p^{\mathrm{T}} x \geqslant y\eta^{\mathrm{T}} e \tag{19.1}$$

for all x, λ and η, u which satisfy the constraints and

$$p^{\mathrm{T}} x^* = y\eta^{*\mathrm{T}} e\,, \tag{19.2}$$

at an optimum. In other words,

$$C(y, p) = y\eta^{*\mathrm{T}} e = p^{\mathrm{T}} x^* \tag{19.3}$$

is the required (minimizing) cost function, which varies with each choice of y and p.

Here we have proceeded from the production function to the cost function but, of course, we could also have proceeded via the opposite course. The latter is only an 'in principle' statement, however, since, as noted in the introduction, many of the inputs and outputs in public sector applications are not easily priced or costed without recourse to arbitrary and ex cathedra procedures and assumptions. Furthermore, the use of price data also has difficulties such as were discussed in connection with Farrell efficiency (in the preceding section) to which we should now like to note further difficulties such as determining whether the data for each DMU refers to list or actual prices and whether allowances have been made for financing considerations, intermingled with production considerations, such as cash discounts, quantity discounts, and so on.

Although we prefer to continue from the production side for reasons such as have just been indicated, there may be cases where prices can also (or alternatively) be used with advantage. We have elsewhere [13] extended 'Shephard's lemma' by dropping the assumption of differentiability and so we now adapt that result for use in the present case as well.

For this purpose we reexpress (19.2) via

$$p^{\mathrm{T}}x^* = y\eta^{*\mathrm{T}}e = y\sum_s \eta_s^* = yc \tag{20.1}$$

where

$$c \equiv \sum_s \eta_s^* , \tag{20.2}$$

so that c is the (total) per unit cost of producing y units of output. Note that we can generally vary the components of p^{T} within some range without altering the x^* values[36]. Thus, if we select one component, say p_i for such variation, we can obtain

$$\Delta p_i x_i^* = y\Delta c \tag{21.2}$$

or

$$x_i^* = y\,\frac{\Delta c}{\Delta p_i} = \frac{\Delta C(y, p)}{\Delta p_i} \tag{21.2}$$

which is Shephard's lemma in discrete form.

The definition of efficiency in (1) has thus enabled us to make contact with this line of work as well. We can also do more. We can open up new possibilities for employing the latter, e.g., in effecting estimates of extremal relations from original data, just as we have done for Farrell efficiency.

To briefly show such possibilities it will suffice to relate Shephard's measure of distance to Farrell efficiency in the context of the models given in Section 2.

[36] Necessary and sufficient conditions as applicable to each component of p^{T} are given in [8], Ch. IX.

This will enable us to deal with this measure in the context of original data. It will also enable us to close this section by returning to the multiple output case.

In Shephard's notation we replace the first portion of (18) with

$$C(y, p) = \min_x \{p^{\mathrm{T}}x : \Psi(y, x) \geqslant 1\} \tag{22.1}$$

where

$$\Psi(y, x) \equiv \frac{1}{\nu^*(y, x)} ,$$

with

$$\nu^*(y, x) = \min \nu ,\ \nu x \in L(y) ,\ \nu \geqslant 0 ,$$

and $L(y)$ is defined as in the second part of (18). This $\Psi(y, x)$ is what Shephard calls a 'distance function'[37].

The above development assumes that the production relations have been determined and hence are available, as in (18.1), for the determination of x. Here, however, we want to regard the latter as data. Therefore, we revert to (8) and write

$$\begin{aligned} &\min \nu \\ &\text{with } \nu X_0 - \sum_{j=1}^{n} X_j\lambda_j \geqslant 0 , \\ &\qquad \sum_{j=1}^{n} Y_j\lambda_j \geqslant Y_0 , \\ &\lambda_j \geqslant 0 ; \quad j = 1, \ldots, n , \end{aligned} \tag{22.2}$$

in which $Y_0 \equiv y$ is now a vector of outputs.

If we regard $z_0^* \geqslant 1$ as an extension of Farrell efficiency to the multiple output case of (9) then we can also regard $\min \nu = \nu^*$ in the above formulation as an extension of Shephard's distance function (and related concepts) for use in evaluating the efficiency of DMU's. As a comparison with (9) makes clear, these z_0^* and ν^* values are complementary to each other. (They are in fact reciprocals. See [13].) In other words, Shephard's distance function is thereby converted into a measure of efficiency with $\nu^* \leqslant 1$ and $\nu^* = 1$ only when the designated DMU is efficient relative to the other observations and/or theoretically prescribed conditions (e.g., via ex cathedra engineering conditions).

In this way we have employed our definition of

[37] Lau in [17], p. 179 – see also Jacobsen in [17], p. 172 – refers to this as a 'gauge function' (its classical denotation in convex set theory) since it does not have the usual properties of a distance function. See, e.g., Appendix A in [8].

efficiency not only to tie together two previously separate strands of economic research, but also to open new possibilities for each of them which include new ways of estimating extremal relations from empirical data.

7. Summary and conclusion

We have now provided a variety of ways of assessing the efficiency of DMU's in public programs in order to improve the planning and control of these activities. This was initiated with a new definition of efficiency, in Section 2, which we have now related to both engineering and economic concepts. We also supplied operational vehicles [38] and interpretations for actual applications. In addition we have introduced a new kind of production function and new methods of securing estimates from empirical data which we have done in ways that have enabled us to bring a variety of economic concepts to bear in new and potentially useful ways. This was done with new models and interpretations, to be sure, but without altering any of the extremization principles which form the essential basis of these concepts.

We can (and will) extend these developments in a variety of additional ways [39]. Here we may best conclude with some comments on possible limitations and alternatives to these approaches.

One limitation may arise because of lack of data availability at individual DMU levels. This is likely to be less of a problem in public sector, as contrasted with private sector, applications. In addition the term DMU can be accorded considerable flexibility in interpretation as when data unavailability may make it desirable to move to school district instead of individual school level and, indeed, it may even be necessary to move to statewide office of education levels when data are not available even at the district level. After all, the objective is to measure the efficiency of resource utilization in whatever combinations are present (loose or tight) in the organizations as well as the technologies utilized [40]. This also suggests a strategy for research in that given the results obtained at one level a systematic basis may be secured for proceeding to other levels by raising pertinent questions and requiring further justification from responsible officials for whatever inefficiencies are uncovered.

Concerning private sector applications, the case for our proposed measure of efficiency begins to weaken to the extent that competition is present. In particular it begins to weaken as soon as freedom for the deployment of resources from one 'industry' to another (perhaps in a removed region) is present. Assessment of such possibilities would involve the introduction of prices, or other weighting devices, for the evaluation of otherwise non-comparable alternatives.

Although our measures are not designed for this sort of application they are designed for public sector programs in which the managers of various DMU's are *not* free to divert resources to other programs merely because they are more profitable – or otherwise more attractive. Our measure is intended to evaluate the accomplishments, or resource conservation possibilities, for every DMU with the resources assigned to it. In golfing terminology it is, so to speak, a measure of 'distance' rather than 'direction' with respect to what has been (and might be) accomplished [41].

[38] I.e., we have supplied both models and computational procedures.

[39] E.g., as already observed the above developments can be extended to functions which are piecewise Cobb–Douglas.

[40] The fact that one must generally include such organization considerations as part of the production function in empirical investigations is in an economic tradition that goes back at least as far as Sune Carlson. See [6]. (In future papers, however, we shall show how the concepts we have introduced here may be extended for use in separating these managerial features in observed data from other aspects of a production technology.)

[41] The term direction should be extended to the plural for the case of multiple outputs. (Here we are concerned with choices of direction which might be effected at the legislative level, including a choice of whether to undertake the program at all.)

References

[1] D.J. Aigner, T. Ameiya, P.J. Poirier, On the estimation of production frontiers: maximum likelihood estimation of the parameters of a discontinuous density function, Internat. Econ. Rev. XVII, (2) (1976) 377–396.

[2] D.J. Aigner and S.F. Chu, On estimating the industry production function, American Economic Review LVIII (1968) 826–839.

[3] R.F. Alioto and J.A. Jungherr, Operational PPBS for Education (Harper and Row, New York, 1971).

[4] K.J. Arrow and F.H. Hahn, General Competitive Analysis (Holden-Day, San Francisco, 1971).

[5] C.R. Bector, Duality in Linear Fractional Programming, Utilitas Mathematica 4 (1973) 155–168.
[6] S. Carlson, A Study on the Pure Theory of Production, (Kelley and Millman, Inc., New York, 1956).
[7] A. Charnes and W.W. Cooper, An explicit general solution in linear fractional programming, Naval Research Logistics Quarterly 20 (3) (1973).
[8] A. Charnes and W.W. Cooper, Management Models and Industrial Applications of Linear Programming (Wiley, New York, 1961).
[9] A. Charnes and W.W. Cooper, Managerial economics – past, present, future, Journal of Enterprise Management (forthcoming).
[10] A. Charnes and W.W. Cooper, Programming with linear fractional functionals, Naval Res. Logist. Quart. 9 (3, 4) (1962) 181–185.
[11] A. Charnes, W.W. Cooper and E. Rhodes, Expositions, interpretations, and extensions of Farrell efficiency measures, Management Sciences Research Group Report, Pittsburgh: Carnegie-Mellon University School of Urban and Public Affairs (1975).
[12] A. Charnes, W.W. Cooper and E. Rhodes, Measuring the efficiency of decision making units with some new production functions and estimation methods, Center for Cybernetic Studies Research Report CCS 276, Austin, TX, University of Texas Center for Cybernetic Studies (1977).
[13] A. Charnes, W.W. Cooper and A. Schinnar, Transforms and approximations in cost and production function relations, Research Report CCS 284, Austin, TX, University of Texas Center for Cybernetic Studies (1977).
[14] Encyclopedia Americana (Encyclopedia Americana Corporation, New York, 1966) 352–353.
[15] M.J. Farrell, The measurement of productive efficiency, J. Roy. Statist. Soc. Ser. A, III (1957) 253–290.
[16] M.J. Farrell and M. Fieldhouse, Estimating efficient production functions under increasing returns to scale, J. Roy. Statist. Soc. Ser. A, II (1962) 252–267.
[17] M.D. Intriligator and D.A. Kendrick, eds., Frontiers of Quantitative Economics, II. North-Holland, Amsterdam (1974).
[18] R. Jaganathan, Duality for nonlinear fractional programs, Z. Operations Res., 17, Physica-Verlag, Wurzburg (197[illegible]) 1–3.
[19] J. Johnston, Statistical Cost Analysis, McGraw-Hill, New York (1960).
[20] K. Lancaster, Consumer Demand, Columbia University Press, New York (1971), 40.
[21] E. Rhodes, Data Envelopment Analysis and Related Approaches for Measuring the Efficiency of Decision Making Units with an Application to Program Follow Through in U.S. Education, Ph.D. thesis, Carnegie-Mellon University School of Urban and Public Affairs, Pittsburgh (1978).
[22] A. Robinson, Non-Standard Analysis (North-Holland, Amsterdam, 1966).
[23] P.A. Samuelson, Prices of Factors and Goods in General Equilibrium, Rev. Econom. Stud. 21 (1953–1954) 1–20.
[24] K. Sato, Production Functions and Aggregation (North-Holland, Amsterdam, 1975).
[25] S. Schaible, Parameter-free Convex Equivalent and Dual Programs of Fractional Programming Problems, Z. Operations Res., 18 (1974) 187–196.
[26] R.W. Shephard, Cost and Production Functions (Princeton University Press, Princeton, 1953).
[27] R.W. Shephard, The Theory of Cost and Production Functions (Princeton University Press, Princeton, 1970).
[28] G.J. Stigler, The Xistence of X-Efficiency, Am. Econom. Rev. (March 1976) 213–216.