WORKING PAPER
ALFRED P. SLOAN SCHOOL OF MANAGEMENT

EFFICIENCY EVALUATION TECHNIQUES:
A CRITICAL COMPARISON AND ASSESSMENT

by

H. David Sherman
Sloan School of Management
Massachusetts Institute of Technology

December 1982

Working Paper 1443-83

MASSACHUSETTS INSTITUTE OF TECHNOLOGY
50 MEMORIAL DRIVE
CAMBRIDGE, MASSACHUSETTS 02139
EFFICIENCY EVALUATION TECHNIQUES: A CRITICAL COMPARISON AND ASSESSMENT

by

H. David Sherman
Sloan School of Management
Massachusetts Institute of Technology

December 1982

Working Paper 1443-83
Efficiency Evaluation Techniques:
A Critical Comparison and Assessment

1. Introduction

This paper addresses the question of how a manager can locate and measure inefficiencies within a set of organizations or decision making units (DMU's) as a basis for planning and implementing changes to improve the efficiency of these DMU's. Specifically, I will compare and evaluate alternative methodologies that have been or may be used to locate such inefficiencies for a particular class of DMU's characterized as follows:

1. They use multiple inputs (resources) jointly to provide or produce a set of multiple outputs (services and/or goods).
2. The production function for the outputs produced cannot be specified in any detail, i.e., there are no available engineered standards as to the amount of various inputs required to produce each output of these DMU's.

The types of DMU's that would be included in this class of organizations include health care organizations (e.g. hospitals), educational institutions, other nonprofit service organizations, and certain types of for profit organizations such as retail bank branches and field service offices for installation and maintenance of equipment. The common element among these organizations is their use of multiple outputs to produce various services and the lack of a clear understanding of the efficient production function, i.e., exactly which and how much of each input is needed to efficiently provide each
unit of service or other output. If a manager knew the efficient input-output relationships, then the determination of which DMU's are inefficient would be easily determined by comparing the actual amount of inputs used to produce their outputs with the amount of inputs that should have been used if the DMU were operating efficiently. In this paper, we specifically consider the managerial problem of locating inefficient DMU's where this efficient input to output relationship is not well understood or specifiable in any detail.

The current growing interest in improving business productivity requires clarification of the aspect of efficiency that is addressed in this paper vis-a-vis the broader concept of productivity. I will emphasize the location of technical inefficiency which relates to the use of physical units of inputs to produce outputs which are also measured in physical units. A DMU that is technically inefficient could produce the same level of outputs with fewer inputs or could produce more outputs with the same level of inputs. Hence, eliminating technical inefficiency can lead to reduced costs without a change in the output's provided or increased outputs without increase in costs. Technical efficiency is the component of productivity that is addressed here. Other components such as price efficiency, i.e., the ability to purchase inputs at the lowest cost, and allocative efficiency, i.e. the use of the optimal mix of inputs to produce the outputs, are not considered in this paper, although references to other dimensions of productivity will be included where appropriate.

Section 2 evaluates several of the techniques that have been widely used to identify inefficiencies of multiple output-multiple input DMU's. This includes simple regression techniques and econometric techniques including simultaneous estimation approaches and flexible functional forms, e.g., the translog function. In addition, the regression based extremal estimation
techniques are considered because in some ways they are more direct in their attempt to evaluate efficiency. In addition, the new and increasingly used linear programming based efficiency evaluation technique developed by Charnes, Cooper, and Rhodes (1979, 1981), referred to as Data Envelopment Analysis (DEA) is considered. Based on comparison of these methodologies, DEA is found to have many attributes which make it superior to the econometric regression techniques. The specific strengths and weaknesses of these techniques are summarized in section 2 and in table 1.

Section 3 illustrates the relative strengths and weaknesses of two of the more comprehensive efficiency evaluation techniques -- DEA and the translog function. These two techniques are emphasized because of their ability to directly address the multiple output-multiple input case when the form of the production function is not specifiable in any detail. By applying these two techniques to a controlled data base, I uncover serious problems with the translog function and get a strong reading of the limits and strengths of using DEA for evaluating efficiency.

Section 4 concludes this paper with an overview of the managerial implication of these results in selection and use of these alternative efficiency measurement techniques, an indication of how managers might employ these techniques and a discussion of areas requiring further research to improve the reliability and usefulness of these techniques.

2. Methodologies for Efficiency Evaluation of Multiple Output-Multiple Input Decision Making Units.

Econometric-Regression techniques have been widely used to estimate production and cost relationships in industries where these relationships are not otherwise specifiable. Rather than cite these numerous applications to nonprofit and for profit organizations, I will make frequent reference to
M. Feldstein's study, *Economic Analysis for Health Service Efficiency*, [1968] because it is thorough in considering a broad spectrum of these techniques and because it is representative of how these techniques are employed in assessing industry efficiency and efficient production relationships within an industry.

The potential problems that confront a manager using econometric-regression techniques to assess efficiency are indicated early on in Feldstein's study [Feldstein, 1968, pp. 33-34]. Feldstein notes that use of econometric-regression based techniques differ from the concepts of Farrell [1962] and Nerlove [1965]. The key difference is that Farrell and Nerlove attempt to locate the set of "best practice" firms among the DMU's under study which would represent the relatively efficient DMU's. Compared with these identified, relatively efficient "best practice" DMU's, the other DMU's can be said to be relatively inefficient. In contrast, the econometric-regression techniques attempt to estimate a mean or maximum likelihood relationship (and not the most efficient or "best practice" relationships). Econometric techniques will only indicate an efficient production relationship when all the units (DMU's) in the study are themselves efficient.

For example, if a set of DMU's with output $y$ and input $x$ were to be evaluated to locate the relatively more and less efficient relationships, the results using regression analysis might appear as in line $ab$ in figure 1, (page 7). This represents the mean relationship of $y = f(x)$ using least square estimation and the slope would represent the amount of input $x$ needed to provide a unit of $y$ on average for the population of DMU's in this study. The question that must be addressed is what information about efficiency is available from line $ab$. For example, does the slope of $ab$ represent the efficient amount of $x$ needed per unit of $y$. One might also question whether a non-linear relationship exists and refit the function with a non-linear function term such as $y = f(ln(x))$ [not shown in Figure 1].
The alternative "best practice" set approach suggested by Farrell [1957] would be to locate the set of most efficient units and consider that to be the set of efficient relationships. In figure 1, the "best practice" set is defined by the line joining observation points C, D, E, and F, each of which represent an observed DMU.

If one could identify this "best practice" set, the manager would be able to draw conclusions such as the following:

1. All DMU's operating below line CDEF such as the DMU's represented by points H and G are relatively inefficient.
2. DMU H should be able to produce as much output y as DMU D with its current input level or it should be able to produce its current output level of y while reducing its inputs to the level of DMU C.

Hence, some insight as to which DMU's are inefficient and the magnitude of inefficiency compared to the "best practice" set are available. The manager would not, however, know from Figure 1 whether DMU's C, D, E, and F are absolutely efficient unless the "true" efficient production function was known. (Recall, that I am interested in efficiency evaluation where the production function is not known. When the efficient production function is known, more direct ways of evaluating efficiency would be available.)

Now consider what information about efficiency is available from the regression analysis that resulted in line ab in figure 1. Line ab represents a mean relationship and therefore is a combination of efficient and inefficient units. One approach to evaluate efficiency of DMU's with regression techniques is to consider those units below the line ab to be potentially inefficient because they obtain less output per unit of input than those above the line. This approach as illustrated in Figure 1 would erroneously exclude H as a potentially inefficient DMU, it would properly capture G as inefficient and would erroneously consider F to be relatively
inefficient. In addition, if slope of line ab were considered to represent the efficient output input relationship for pricing or other managerial purposes, it may misrepresent the true relationship. For example, the slope of a regression line based only on the most efficient DMU's might look like line ln which suggests a different (in this case higher) amount of inputs x needed to produce a unit of y than is estimated for line ab.

In spite of the potential problems associated with use of the econometric regression techniques described above, these techniques have been widely used. Feldstein [1968] specifically opted for these techniques over the Farrell [1962] Concepts for two reasons: 1) the estimation techniques of Farrell and Nerlove were more complicated than the econometric-regression techniques and 2) Farrell's approach only allowed use of a small fraction of the total observations.

Since Feldstein's study, a new technique referred to as Data Envelopment Analysis (DEA) based on Farrell's concepts was developed which attempts to locate the "best practice" set of DMU's which specifically use multiple inputs to produce multiple outputs. This methodology provides an explicit efficiency evaluation technique which can be readily applied with current computer technology and which will now be compared with standard econometric-regression techniques. Other more sophisticated econometric techniques attempt to remedy the problems noted in our discussion of Figure 1. Specifically, I will also consider the translog (flexible functional form) function which has been developed to handle the multiple output/input case without the need to specify a functional form and I will consider extremal type regression that attempt to estimate the "best practice" production relationships with respect to efficiency assessment capabilities.
"best practice" set of DMU's
least squares estimate
\[ y = f(x) \]
I have summarized the comparisons of these techniques in table 1 which list the efficiency evaluation methodology on the left side of each rows and the criteria for assessing the methodology on the top of each column. Each box in the matrix in table 1 indicates the capability of each methodology with respect to the criteria in that column and the (+) and (-) indicates whether this is a positive or negative attribute of the methodology.

Criteria to Assess the Alternative Efficiency Evaluation Methodologies

A set of general criteria have been identified which I believe to be common issues in evaluating an organizations relative efficiency. These criteria embrace the clear trends in econometrics which acknowledge the need to develop techniques that accommodate multiple outputs and inputs, and the need to address industries with unspecified production relationships (see for example [12]). In addition, a review of other efficiency evaluation studies including Sherman [1981], CCR [1981], Gold [1982], Bessent & Bessent [1979], and Lewin and Morey [1981] tempered by extensive involvement and discussion with management of many of the organizations in these studies have allowed me to expand this group of criteria that seem to be reasonably comprehensive in locating issues that these methodologies need to address.

The criteria are briefly described below.

I.

I. The methodology should explicitly incorporate multiple outputs and inputs.

This requirement is basic to the type of evaluation needed for a broad group of institutions including many nonprofit organizations, e.g., hospitals, educational institutions, public service organizations, as well as certain primarily service based for
profit organizations, e.g., branch banking and multioffice consulting firms. A methodology that is unable to handle multiple outputs and inputs would either require that these inputs and outputs be evaluated one-at-a-time thereby ignoring their joint nature or it may require collapsing the outputs and/or inputs into a single unit of measure which can only be done where these have common units and an objectively determined weighting system. For example, hospitals treat many types of patient diagnoses. If we cannot explicitly include each diagnostic category as an output, we would have to collapse these outputs into one output measure via a system that puts all patient case into common units and appropriately weights the individual diagnoses to reflect the variations in severity and level of care. In many industries like health care (see Sherman [1981]), the needed objective weighting schemes are not available.

II. The methodology should accommodate DMU's for which the input-output relationships, i.e., the functional form are not specifiable. The types of institutions which are to be evaluated do not have clear and readily specified production functions. For example the efficient amount of inputs to care for a patient with a particular illness or to train a student are not known. Hence, any evaluation tool that requires specification of a functional form will have limited value.

III. The methodology should help identify inefficient behavior versus more efficient behavior rather than describe the behavior of all efficient and inefficient DMU's combined. The behavior of efficient and inefficient units averaged together does not provide the types of insights needed to help determine the level of inefficiency within a DMU and the degree of improvement possible if the inefficiency is remedied.
IV. The methodology should clearly identify the inefficient DMU's versus the more efficient DMU's rather than rely on arbitrary or subjective cutoff points for separating efficient and inefficient DMU's. This is of key importance in utilizing the results, as the manager of a DMU which has been subjectively categorized as inefficient may argue with the arbitrary cutoff and thereby weaken the ability to use the methodology to improve efficiency.

V. The methodology should go beyond the designation of efficient versus inefficient and provide insights as to the magnitude of the inefficiency as well as the source of the inefficiency. This quality would help translate the results into managerial action.

VI. The methodology should provide insights to help define efficient behavior of a DMU in its industry beyond those included in the data set via statistical inference or other means. This quality would extend the results beyond the identification of the inefficient units in the data set to provide some information about how an efficient DMU is operated. This criteria is most relevant where it is difficult to include the entire population in the efficiency evaluation. In addition, satisfying this criteria would help to determine the as yet unknown efficient production function of the industry.

Based on these six criteria, a set of widely used methodologies are evaluated along with certain more recently developed techniques which also represent useful attempts to evaluate efficiency and, therefore, seem appropriate for consideration in this paper.
Evaluation of Alternative Efficiency Evaluation Methodologies based on the established criteria.

As summarized in table 1, DEA is the only methodology that meets five of the six criteria. Extremal regression techniques meet four of the criteria (in part) and the translog function (flexible functional form) meets only two of the criteria. Extremal techniques also appear to be inadequate as currently developed for two reasons: they have not yet been extended to the multiple output-multiple input case, and 2) they require specification of the functional form (see [Lovell and Sickles 1980] and [Forsund et al. 1980]). These limitations represent two characteristics of primary importance for the types of evaluation of concern in this paper. The other regression based techniques meet none or only one of the criteria and are, therefore, clearly inappropriate as efficiency evaluation techniques.

DEA and the translog function will be evaluated in greater depth in section 3 because DEA appears to meet all but one of these criteria and the translog function meets the two primary criteria and has been used with increasing frequency to address the efficiency evaluation of the multiple output-input firms of concern in this paper.

Having suggested the conclusions of this inquiry into the relative value of these methodologies, the following section discusses these methodologies and the basis for the assessment of each technique with respect to the six criteria described above.

A. Simple Regression

Simple regression refers to the use of single equation models of the form:

\[ y = f(x_i) \quad i = 1 \ldots n. \]  

(1)
Examples of this are Feldsteins estimation of production functions of hospitals where $y$ is the sum of patients treated weighted by the relative actual cost (not efficient cost) of each type of patient and $x_i$ represents the various inputs and other independent variables such as location and size. Typically, alternative formulation of (1) might be attempted to determine if a linear or non-linear form result in a better fit with the data. This approach can only accommodate a single output variable and immediately requires specification of a functional form thereby failing Criteria I and II. The need to use a least squares technique results in a mean or maximum likelihood relationship which results in a central tendency rather than efficient set of relationships thereby failing criteria III. No objective data are provided about which DMU's are inefficient which results in failure to meet criteria IV and V. This is illustrated by the previous discussions of figure 1 where one could not accurately segregate efficient DMU's like F from inefficient DMU's like G and H. While the simple regression technique does provide a basis for statistical inference about the entire population of data, this inference would not describe efficient behavior. Rather, such inference would be useful only for pure prediction about the behavior assuming the level of inefficiency is held constance in the population as is discussed in Sherman [1981] and CCR [1981]. Hence, criteria VI is also not satisfied.

The problems associated with simple regression may be illustrated by considering typical ways it has been used. Based on the simple regression results, one might conclude that there are constant return to scale if the linear form results in a better fit or economies of scale if the non-linear form results in a better fit. This would overlook the questions of whether there are increasing returns with respect to certain outputs and decreasing returns with respect to other outputs. In addition, what may appear as
returns to scale may really be due to inefficiencies unrelated to scale among the larger or smaller DMU's in the data set. Other conclusions that might be reached are the marginal cost of added outputs and rates of substitution among the inputs. Again, such results are based on the coefficients of the least square regression which reflect mean rather than efficient relationships. Hence, these types of insights may be useful for prediction assuming the level of inefficiency is constant but they provide little help in determining the efficient rates of substitution and efficient marginal costs. Finally, the relative weights often used to collapse multiple outputs into a single output measure are often unavailable or subjective. To the extend that they are subjective or based on average rather than efficient relationships, these weights can bias the results.

B. Extremal Regression approaches to estimate Efficient Production Functions

A direction of research has evolved to address ways of estimating the efficient production function rather than the central tendency function; a key problem encountered with the simple regression using least squares. These approaches are reviewed in detail by Forsund, Lovell, and Schmidt [1980] and examples of these techniques are illustrated in Figure 2 and discussed below. The ordinary least squares estimate for a single independent variable is obtained for such a regression in the form

$$\min f(a, b) = \sum_{i=1}^{m} (y_i - a - bx_i)^2$$

(2)

where $y_i$ and $x_i$ are the values of the $i^{th}$ observation. $x_i$ is the independent variable and $y_i$ is the dependent variable. The estimated values of $a$ and $b$ are otherwise unconstrained. The resulting estimated relationship is represented by the broken line LM shown at the bottom of Figure 2 when
$y = a + bx$ constrained by the condition that $y_i \leq a + bx_i$, $i = 1, \ldots, m$

$y = a + q + bx$

ordinary least squares estimate of $y = a + bx$

Figure 2
fitted to the observations represented by dots (•) each of which has a coordinate \((x_1, y_1)\). Thus LM is meant to reflect a least squares estimate of \(y = a + bx\) that results from minimizing the sum of the squared deviations as in (2).

Aigner and Chu [1968], however, argue that the economic theory of production requires that the above estimation model must be further constrained to satisfy the following \(m\) constraints.

\[
a + bx_i > y_i, \quad i = 1, \ldots, m
\]  

(3)
since the true production function must in each case be able to yield at least the observed \(y_i\) for the amount of input \(x_i\) which was used. This result is approximately reflected by the solid line NO in figure 2. Evidently, it runs along the upper boundary of the \(y_i\) values at each \(x = x_i\) for the independent variable.

The approach of Aigner and Chu is not the only possibility and another approach is discussed by Forsund, Lovell, and Schmidt [1980] in the course of their review of various approaches (including (3)) to estimate an efficient production frontier. They proceed to develop a regression estimate with one-sided error terms by first estimating the input-output relationship in an ordinary manner as illustrated by line LM in Figure 2. Then holding the slope constant, they raise the intercept point L to the lowest value that will raise the regression line to a position in which it is at or above all the observed \(y_i\) values. This results in moving the least square regression line up some distance \(q\) as reflected by the broken line EF in Figure 2.2 which is parallel to line LM and may be represented as

\[
y = a + q + bx \tag{4}
\]

where \(a + q = \hat{a}\) represents a new intercept value and \(b\) is the already estimated least squares value of the slope.
The deviation of each \( y_1 \) value below the line \( EF \) represented by (4) may then be used as a measure of inefficiency for this output at its associated input value. The value, \( q \), may also be regarded as a measure of overall efficiency gain relative to the least square line MN (see Figure 2). The adjustment that was effected from \( a \) to \( a + q \) also carried with it strong assumptions of independence. Note, in particular, that \( q \) is an unknown to be estimated and hence cannot be treated in a manner analogous to replacing the value of \( a \) by a known value of the intercept constant in ordinary least squares analysis. Work is also underway on extending these ideas for estimating extremal relations to multiple output-input conditions. Lovell and Sickles [1981], for example, combine these ideas with those of flexible functional forms to try to estimate efficiency.

With respect to the criteria we have established, extremal regressions are not yet adequately developed to deal with the multiple output-input case, and they require that a functional form and the relationships among certain variables be specified or assumed; hence, these fail criteria I and II. They seek the efficient relationships and help locate inefficient units thereby satisfying criteria III and IV. While they satisfy criteria V in their ability to suggest the magnitude of the inefficiency present, they do not provide clear insight into the source of the inefficiency and, therefore, meet criteria V only in part. Statistical inference to the entire population is available and consequently extremal approaches would satisfy criteria VI.

Forsund et. al. [1980] also describe other problems with alternative extremal regression approaches. As these problems are resolved, this approach may prove to be a highly effective technique for efficiency evaluation of multiple output input firms.
C. Simultaneous and Recursive econometric estimation techniques

The key advantages of a multiple equation system over the simple regression techniques with respect to the established criteria is the ability to explicitly include multiple outputs and inputs. This is done by developing a separate equation for each output and by simultaneously or recursively estimating the coefficients. While this attribute satisfies the first criteria, these multiple equation systems require explicit information about the functional forms. In the recursive type of model, such as is used by Feldstein [1968] for hospital production function estimation, there is also the need to know how the equations relate to each other. Hence, these systems do not meet criteria II because of the extensive insight into the production relationships that is needed a priori. The multiple equation econometric methodologies also use mean or central tendency estimation techniques which, like the simple regression case, do not help seek out efficient relationships, do not separate out the inefficient DMU's, and consequently do not provide data on the magnitude of the inefficiency. Hence, criteria III, IV, and V are not satisfied. With respect to criteria VI, statistical inference from multiple equation models is only possible assuming the degree of inefficiency is constant throughout the population and they don't indicate what efficient behavior is in the population unless each member of that population is coincidentally operating efficiently.

D. Translog Function

The translog function is an increasingly used estimation technique from the class of flexible functional form methodologies. The translog function explicitly comprehends multiple outputs and inputs and specifically does not
require that a particular functional form be designated \textit{a priori}. An example of how this translog function would appear for two outputs $y_1, y_2$ and two inputs $x_1, x_2$ is as follows:

$$\log y_1 = \log a_{00} + a_{10} \log x_1 + a_{20} \log x_2 + b_{20} \log y_2 + a_{11} \log^2 x_1 + a_{22} \log^2 x_2 + b_{22} \log^2 y_2 + a_{12} \log x_1 \log x_2 + a_{13} \log x_1 \log y_2 + a_{14} \log x_2 \log y_2, \text{etc.}$$

Least squares analysis may be used, to determine the constraints $a_{00}, \ldots, a_{14}$. Resulting significance tests now extend to determining whether $y_1$ and $y_2$ are complements or substitutes while $a_{10}$ and $a_{20}$ are available for determining the nature of any scale effects that may be present and the coefficients associated with the higher order terms assist in identifying additional properties such as second order effects, substitution effects between inputs, or between outputs, and so on.

This approach can be extended to deal with many outputs and inputs and clearly satisfies criteria I and II. Nevertheless, it also fails to meet the other criteria because it is a central tendency estimation technique and therefore does not help to identify efficient behavior. Consequently, conclusions from studies using translog function like Bothwell and Cooley [1978] about scale economies and rates of substitution among inputs and outputs do not necessarily reflect efficient relationships.

E. Data Envelopment Analysis (DEA) - a New Methodology for Measurement of Relative Technical Efficiency

Data Envelopment Analysis represents a new methodology for efficiency evaluation which promises to circumvent many of the difficulties encountered
with the estimation techniques discussed above. Charnes, Cooper and Rhodes
[CCR 1979, 1981] generalized the usual input to output ratio measure of
technical efficiency to the conditions for multiple inputs and multiple
outputs in terms of a fractional linear program with fractional constraints
which is represented in the following model:

Objective:

$$\max h_o = \frac{\sum_{r=1}^{s} u_r y_{ro}}{\sum_{r=1}^{m} \sum_{i=1}^{n} v_i x_{ij}}$$

where $o$ is the Decision-Making Unit (DMU) being
evaluated in the set of $j = 1, ..., n$ DMU's.

Constraints:

Less than

Unity : $1 > \frac{\sum_{r=1}^{s} u_r y_{rj}}{\sum_{i=1}^{m} v_i x_{ij}}$ ; $j = 1, ..., n$  \hspace{1cm} (6)

Positivity Constraints : $0 < u_r ; r = 1, ..., s$

Data:

Outputs: $Y_{rj} =$ observed amount of $r^{th}$ output for the $j^{th}$ DMU

Inputs: $x_{ij} =$ observed amount of $i^{th}$ input for the $j^{th}$ DMU

The $u_r$ and $v_i$ values are determined objectively from the data in terms
of the above model. In solving for $\max h_o = h_o^*$, the efficiency of

DMU $j = o$ is evaluated such that

$$h_o^* = 1 \text{ if and only if } \text{DMU } j = o \text{ is efficient}$$

relative to the other DMU's that are

represented in the constraints, and

$$h_o^* \leq 1 \text{ if and only if } \text{DMU } j = o \text{ is}$$

relatively inefficient relative to other

DMU's that are represented in the constraints.
The above formulation provides a model which generalizes the usual single output-single input measure of efficiency (as in engineering or physics) so that it also embraces multiple output-multiple input situations. Applied to empirical data, it provides a new way of estimating extremal relations as well as measuring the relative technical efficiency of multiple output-multiple input type DMU's where the production relationships are not readily specified in detail. Hence, criterion I and II are met with DEA.

Applying DEA to a set of DMU's results in an efficiency rating for each DMU of 1 (efficient) or less than 1 (relatively inefficient). These ratings, however, represent relative efficiencies obtained from a piecewise linear production possibility frontier comprised of the most efficient units in the set of \( j = 1, \ldots, n \) DMU's. This "best practices set" is theoretically consistent with microeconomic analysis. It differs from the efficient production function and the derived efficient isoquant analysis found in economics only to the extent that these efficient units are relatively, rather than absolutely efficient.¹

---

¹The above formulation, (6), involves a nonlinear-nonconvex programming problem. As is shown in CCR [1979] however, it may be replaced by dual linear programming problems as follows:

\[
\begin{align*}
\text{Max } h_0' & = \sum_{r=1}^{s} u_r y_{ro} \\
\text{Subject to } & \\
0 & > \sum_{r=1}^{s} u_r y_{rj} - \sum_{i=1}^{m} \omega_i x_{ij}; \quad j = 1, \ldots, n \\
1 & = \sum_{i=1}^{m} \omega_i x_{i0} \\
0 & < m_r, l_i; \quad r = 1, \ldots, s \\
& \quad i = 1, \ldots, m
\end{align*}
\]
In locating efficient and inefficient DMU's, DEA satisfied criteria III and IV. In addition, interpretation of the DEA results indicate the magnitude of inefficiency present as well as the output and input adjustments needed to make an inefficient DMU efficient as is illustrated in Sherman [1981] for example. Hence, criteria V is also satisfied. Finally, with respect to criteria VI, DEA is not currently capable of defining the efficient industry production relationship by extending the results beyond the observed DMU's.

It is not surprising that a methodology like DEA designed for efficiency evaluation might be more powerful than adaptation of methodologies like econometric regression techniques for evaluating efficiency. In the next section, DEA and the translog function are examined more closely in an application to further clarify the relative merits of the two methodologies that appear to meet the two key criteria for the type of efficiency assessment that is sought, i.e., the ability to evaluate DMU's with multiple outputs and inputs in the absence of well-defined production function.

3. Application of DEA and the Translog function to assess efficiency in a controlled set of DMU's

Numerous studies have used the Translog function to provide a structure to estimate production and cost relationships among organizations (see for example Bothwell & Cooley [1978]). Other studies have used DEA for evaluating efficiency of organizations (see for example [Bessent & Bessent] and [Lewin & Morey]). Rather than follow along these lines, I will use an artificial data set to strengthen the ability to compare DEA and the Translog function capabilities in assessing efficiency. This is necessary because the typical DMU's for which these methodologies are used do not have known production functions. Consequently, in a real data set there is no objective way to determine if truly inefficient or efficient units are correctly identified as such with whichever methodology is being employed.
Sherman [1981] created an artificial data set by arbitrarily defining the efficient production relationships for a hypothetical hospital that used three inputs (full time equivalents of personnel, supply dollars and beddays available) to produce three outputs (severe patient care, regular (less resource intensive) patient care, and training of interns and residents). These efficient relationships applied to hospitals of any size so constant returns to scale were present. Seven of the hospitals (H1 - H7) were developed as efficient hospitals because they used exactly the amount of inputs to produce their output levels as was required by the defined efficient production relationships. It was also assumed that all hospitals paid the same price for each unit of each input so price efficiency was not a cause of different levels of efficiency. Eight of the hospitals (H8 - H15) were designed to be inefficient in that they each used more inputs than was required by the efficient production function to produce their respective levels of outputs. The details of the development of this data base are reported in Sherman [1981] and summarized in exhibits 1, 2 and 3. The motive in using this data base here is to gain insights about how well DEA and the translog function perform in application to this relatively simple example. If problems arise here, they would not be likely to diminish in application to a real data base where there may be more variables and observations and where there would be no objective knowledge about which DMU's were efficient and inefficient.

**Translog Cost function**

The Translog function has been recently used by Bothwell and Cooley [1978] to study the presence of scale economies (with positive results) in the case of Health Maintenance Organizations. For a variety of reasons, it is easier to study returns to scale with cost functions rather than production
functions, and this, in any case, is the way many studies in the area have approached this topic. In addition, this approach is of interest in areas like hospitals where the primary concern has been cost behavior rather than input output relationships addressed via production functions. Bothwell and Cooley proceed to estimate the following translog cost function:

$$\log C = a_0 + \sum_{i=1}^{m} \alpha_i \log y_i + \sum_{j=1}^{n} \beta_j \log p_j$$

$$+ \frac{1}{2} \sum_{i=1}^{m} \sum_{e=1}^{m} \delta_{ie} \log y_i \log y_e$$

$$+ \frac{1}{2} \sum_{j=1}^{n} \sum_{k=1}^{n} \gamma_{jk} \log p_j \log p_k$$

$$+ \sum_{i=1}^{m} \sum_{j=1}^{n} \rho_{ij} \log y_i \log p_j + \cdots \quad (8)$$

where

- $C$ = total cost
- $p_j$ = the price per unit of the $j^{th}$ input
- $y_i$ = the amount of the $i^{th}$ output.

We shall use this approach and estimate the following translog cost functions for the 15 artificial hospitals.

$$\ln C = \alpha_0 + \alpha_1 \ln S + \alpha_2 \ln R + \alpha_3 \ln T$$

$$+ \beta_{11} (\ln S)^2 + \beta_{22} (\ln R)^2 + \beta_{33} (\ln T)^2$$

$$+ \delta_{12} \ln S \ln R + \delta_{13} \ln S \ln T + \delta_{23} \ln R \ln T \quad (9)$$

where

- $C$ = Total Cost per year
- $R$ = Number of regular patients treated in one year.
- $S$ = Number of severe patients treated in one year.
- $T$ = Number of interns and residents trained in one year.
It will be seen that the terms involving input prices have been omitted. By construction in the data, all input prices are constant and also were assumed to have no effect on output levels. This is to say that omission of both the pure price terms and the terms allowing for price and output interactions is entirely justified. Also, however, the term involving \((\ln T)^2\) which was found to be highly colinear with other independent variables will also be omitted which means that there is some possibility of bias in the coefficient estimates.

By construction, the cost function applies in each hospital. Hence, only one translog function from the data needs to be estimated which yields

\[
\ln C = 23.2 + .825 \ln S - 3.044 \ln R \\
- 2.459 \ln T - .069 (\ln S)^2 + .161 (\ln R)^2 \\
+ .031 \ln S \ln R + .147 \ln S \ln T + .173 \ln R \ln T
\]  

(10)

with

\[
R^2 \text{ (unadjusted)} = 0.992
\]

\[
S_c \text{ (Standard deviation)} = 0.026
\]

Following the type of analysis of Bothwell and Cooley, the analysis is broken into parts in order to ascertain whether increasing or decreasing returns to scale are present (a) in any of the outputs and (b) in their overall, combined effects.

The fact that the translog is a flexible functional form allows for the possible presence of increasing returns to scale in some output range and decreasing returns to scale in other output ranges. For the moment, however, these possibilities are set aside and it is assumed that the properties are global. Hence it is assumed that the function can be evaluated at any positive level of output with the result that the properties noted at their output level will apply over the entire relevant ranges of the function.
<table>
<thead>
<tr>
<th>Step 1</th>
<th>Step 2</th>
<th>Step 3</th>
<th>Step 4</th>
<th>Step 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
</tbody>
</table>

**Multiple Output-Multiple Input Decision Making Units**

ATTRIBUTES OF METHODOLOGIES TO EVALUATE EFFICIENCY OF

**Table 1**
Following Bothwell and Cooley I chose for the prescribed output level the base of the natural logarithm system—i.e., I evaluate the properties of C at \( S = R = T = e = 2.718 \ldots \), approximately. Higher order terms involving these variables then become unity and evaluations for the presence of the indicative scale properties becomes one of adding exponents as described in Bothwell and Cooley.

**Case i. Partial Scale Economies**

(i.i) Severe Patients: Increasing returns to scale are present

\[
\frac{\partial \ln C}{\partial \ln S} = .825 = 2(.069) + .031 + .173 \quad 0.87 < 1 \quad (11)
\]

(i.ii) Regular Patients: Increasing returns to scale are present

\[
\frac{\partial \ln C}{\partial \ln R} = -3.044 + 2(.161) + .031 + .173 \quad -0.47 < 1 \quad (12)
\]

(i.iii) Training: Increasing returns to scale are present

\[
\frac{\partial \ln C}{\partial \ln T} = -2.452 + .147 - .173 \quad -2.13 < 1. \quad (13)
\]

In addition as developed in Bothwell and Cooley

\[
\frac{\partial C}{\partial R} = -0.47 \frac{C}{R} \quad (14)
\]

\[
\frac{\partial C}{\partial T} = -2.13 \frac{C}{T}
\]

which means that marginal costs for regular patients and training are not only falling but are also negative with, consequently, falling **total** costs as well as these outputs **increase**.
Case II. Overall Scale Economies

Following Bothwell and Colley's development one can determine the existence of overall scale economies which they show to be

\[
SE = \sum_{i=1}^{m} \frac{\ln C_i}{\ln Y_i} 
\]

and in this case from (i.i), (i.ii), and (i.iii) above:

\[
SE = .87 + .047 - 2.73 = -2.13 < 1. 
\]

This suggests that there are overall increasing returns to scale which, of course, we know are not really present.

Additional Analysis of Scale Economies

The above analysis proceeded on the assumption that an evaluation conducted at the point \( e = R, S, T \) applied globally as well. This is evidently not the case for our function (10) if one assumes that all coefficients are statistically significant. Hence dropping the assumption of globality one can gain additional insight by using inverse logs on \( C \) in order to write our total cost function:

\[
C = \text{antilog (23.2)} \left( 0.825 - 0.069(\ln S) \right) \frac{\text{antilog (23.2)} \left( 0.825 - 0.069(\ln S) \right)}{R^{3.044 - 0.161(\ln R) - 0.034(\ln S)} T^{2.459 - 0.147(\ln S) - 0.173(\ln R)}} 
\]

The expression for \( C = C(y;p) \) where \( y \) represents a vector of outputs and \( p \) a vector of input prices, is missing the terms for input prices. These, however, are all constant and hence only enter into the level of \( C \) and do not otherwise affect its qualitative behavior. Our interest is only in the latter, however, and only in the broad general terms which are summarized by examining the cost function near the origin and within the range of the hypothetical hospital data as follows:
Near the origin, (columns a, b, and c above), total cost (C) is falling as all outputs, R, S, and T increase, i.e., as R, S, and T simultaneously increase from 1 to 2 to 4 the total costs decrease from $11.9 Billion to $595 Million to $47.6 Million. Within the range of the hospital outputs in the data set represented by columns d and e above, the total cost increases. Moreover, outputs in column e are twice the column d outputs while the cost increases disproportionately reflecting sharply decreasing returns to scale within this range.

A more precise characterization would involve complex consideration of the nonlinear interactions between the outputs which are exhibited in (17). Use of the translog function estimate results initially in a broad qualitative characterization of falling total costs followed by decreasing returns to scale in that order as output expands. This is true for each output separately as well as all outputs considered together. Yet none of these properties of scale economies are present in any hospital or in the aggregate of hospitals either since, by construction, the same constant return to scale function was applicable in each case. This all occurs in almost textbook
fashion so that, accompanied by tests of statistical significance, the result would almost surely be accepted as conforming to the requirements of both statistics and economic theory. Such conclusions are quite possible even though these results are known to be false.

A reader may, at this point wonder what has gone wrong or may believe that the data were deliberately arranged to achieve these results. These results were as surprising to this researcher (at first sight) as they may appear to a doubting reader. Even if the charge of deliberate prearrangement is true, however, the result nevertheless admits the possible occurrence of such situations in practice where no charges of prearranged results can be levelled.

The source of what has gone wrong is evidently the presence of inefficiencies with a resulting mismatch between the efficiency requirements of economic theory and the statistical methods employed. The optimizations used in the statistical methods that are commonly employed (e.g., least squares estimation techniques) allow the inefficient units to be considered simultaneously with the efficient ones with a resulting function of best fit that contains influences from both types of behavior. Note, therefore, that economies can come from reduction of inefficiencies—i.e., technical inefficiencies—as well as changing scales. This is already ambiguous, to say the least, and may actually produce erroneous inferences (as above) even with respect to broad qualitative properties like increasing and decreasing returns to scale. The point to bear in mind is that such properties as increasing and decreasing returns to scale can have unambiguous meanings only when technical efficiency have been achieved. A fortiori the same applies to elasticities of substitution and other coefficient estimates which have played such prominent roles in these regression-type approaches.
### Table 1

<table>
<thead>
<tr>
<th>Efficient DMU's</th>
<th>DEA Efficiency Rating (E)</th>
<th>Efficiency Reference Set</th>
</tr>
</thead>
<tbody>
<tr>
<td>H1</td>
<td>1.0</td>
<td></td>
</tr>
<tr>
<td>H2</td>
<td>1.0</td>
<td></td>
</tr>
<tr>
<td>H3</td>
<td>1.0</td>
<td></td>
</tr>
<tr>
<td>H4</td>
<td>1.0</td>
<td></td>
</tr>
<tr>
<td>H5</td>
<td>1.0</td>
<td></td>
</tr>
<tr>
<td>H6</td>
<td>1.0</td>
<td></td>
</tr>
<tr>
<td>H7</td>
<td>1.0</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Inefficient DMU's</th>
<th>DEA Efficiency Rating (E)</th>
<th>Efficiency Reference Set</th>
</tr>
</thead>
<tbody>
<tr>
<td>H8</td>
<td>0.99</td>
<td>H4</td>
</tr>
<tr>
<td>H9</td>
<td>0.98</td>
<td>H1, H2, H6</td>
</tr>
<tr>
<td>H10</td>
<td>1.0</td>
<td>H1, H2, H6</td>
</tr>
<tr>
<td>H11</td>
<td>0.85</td>
<td>H4, H7</td>
</tr>
<tr>
<td>H12</td>
<td>0.99</td>
<td>H1, H4, H6</td>
</tr>
<tr>
<td>H13</td>
<td>1.0</td>
<td></td>
</tr>
<tr>
<td>H14</td>
<td>0.99</td>
<td>H1, H4, H6</td>
</tr>
<tr>
<td>H15</td>
<td>0.87</td>
<td>H4, H6, H7</td>
</tr>
</tbody>
</table>
DEA Applied to Artificial Data Set

DEA was applied to the same 15 artificial hospitals using the linear form described in CCR [1979]. The actual application is illustrated in detail in Sherman [1982]. The results are summarized in table 2 where the efficiency rating of $E = 1.0$ suggests relative efficiency and $E < 1.0$ suggests the presence of inefficiencies. In addition, for each inefficient DMU (in this case hospital unit), the efficient units against which the inefficient unit is most clearly found to be inefficient are identified and these are indicated in table 2 and referred to as the efficiency reference set.

DEA has accurately identified the efficient units ($E = 1$) and it accurately located 6 of the inefficient units ($E < 1$). DEA did not however identify two inefficient units H10 and H13 as inefficient i.e. they were rated $E = 1.0$.

The technical reason that two inefficient units were not identified is discussed in Sherman [1982]. It is sufficient to note that any DMU rated as efficient is only relatively efficient and that DEA will not necessarily be powerful enough to locate all the relatively efficient units.

The power of DEA is in its location of inefficient DMU's. When such DMU's are located, one can conclude that other observed DMU's have operated more efficiently. This suggests that the inefficient DMU can become more efficient by lowering costs or increasing output. Furthermore this improvement might be effected by adopting the operating technique found in the relatively efficient units, e.g., those of the efficiency reference set. It is also possible to identify the magnitude of change in input or output levels that would make the DMU efficient as is illustrated in Sherman [1981, 1982].

In summary, DEA has the ability to locate relatively inefficient DMU's in multiple output–multiple input situations where the production function is not specifiable. It provides information about the magnitude of the inefficiency
present and the potential sources of that inefficiency which may help managers
to reallocate resources and transfer operating techniques to improve
efficiency of their inefficient DMU's. This methodology does not, however,
provide direct access to the true production function. Nevertheless, compared
to the other alternatives described in this study, DEA provides unambiguous
ways of identifying inefficiencies which can be translated into managerial
actions to improve efficiency.

4. Conclusion

I have considered alternative approaches to evaluating efficiency for the
specific care of a multiple output-multiple input DMU where the production
function is not specifiable. Although our examples have tended to emphasize
health care DMU's, this represents a broad group of DMU types as is reflected
in studies of armed forces recruiting offices (Lewin & Morey [1981]), public
education programs (CCR [1981]), and other non-profit organizations. It also
applies to service type DMU's, like branch banking, insurance companies,
customer service organizations, etc.

The findings suggest that existing econometric approaches which may be
well suited for pure prediction purposes do not perform well in evaluating
efficiency. The translog function exhibited serious problems in a very simple
application which are not likely to be diminished in more complex situations.
The extremal regression techniques would appear to have some promise when the
multiple output-multiple input case is adequately addressed, though Forsund
et. al. [1980] indicate that the strong assumption they are likely to require
may weaken this approach, particularly when we have inadequate knowledge of
the production function to verify these assumptions.
DEA appears to be most promising in that it has been found to be reliable in its location of inefficient units and its information about the magnitude of the inefficiency. In addition, it provides clear indication of the potential source of the inefficiency which allows the results to be translated into management action. Further research is needed to better understand the degree to which inefficient units may escape identification via DEA.

Finally, it may be possible to combine DEA and regression technique to help identify the efficient frontier production function. For example, DEA might be used to select out the inefficient DMU's it identifies so that the econometric-regression approaches can be applied to a set of observations with fewer inefficient DMU's. This might result in an estimated (least squares) production function which is closer to the efficient production function.

Most directly, this study indicates that greater circumspection is needed in use of econometric techniques for efficiency evaluation and more use and development of DEA would be helpful for efficiency evaluation.
### Exhibit I: Hospital Production Model (Efficiency Model Operational)

<table>
<thead>
<tr>
<th>Patient Type</th>
<th>Training Unit</th>
<th>Treatment Unit</th>
</tr>
</thead>
<tbody>
<tr>
<td>Patient</td>
<td>500 applicants/yr</td>
<td>0.3 FTE/per</td>
</tr>
<tr>
<td>Severe</td>
<td>90 bed-days/per</td>
<td>0.05 FTE/per</td>
</tr>
<tr>
<td>Care</td>
<td>70 bed-days/per</td>
<td>0.04 FTE/per</td>
</tr>
</tbody>
</table>

**Supply & Demand**

- Full-time equivalents
- Bed-days required to efficiently produce one unit

The simulated data base in Exhibit 2 demonstrates the accuracy of the Hospital Production Model to create the expected output and cost relationships as assumed in the model.
**Note:** The table and diagram are not transcribed accurately due to the quality of the image. For a more detailed analysis, higher quality images or clearer scans would be necessary.

---

**Appendix 2**

---

**Exhibit 1**

**Constructed Data Base**

---
(continued next sheet for HL and H15)

<table>
<thead>
<tr>
<th>HL</th>
<th>H15</th>
</tr>
</thead>
<tbody>
<tr>
<td>200 - 00</td>
<td>0.03 - 0.03</td>
</tr>
<tr>
<td>00 - 30</td>
<td>0.05 - 0.05</td>
</tr>
<tr>
<td>30 - 50</td>
<td>0.04 - 0.04</td>
</tr>
</tbody>
</table>

Total Patients Admitted

Supplementary Data

- Teaching hospitals
- General hospitals
- Specialty hospitals
- Other hospitals

Example of construction of data base for hospitals (efficient) and H15 (inefficient)
REFERENCES


