95-B-6

A Survey of Some Recent Developments in Data Envelopment Analysis

W.W. Cooper
and
Kaoru Tone

W.W. Cooper* and Kaoru Tone**

A SURVEY OF SOME RECENT DEVELOPMENTS IN DATA ENVELOPMENT ANALYSIS***

Abstract:

This is a survey of selected recent developments in DEA. Included are (1) new measures for evaluating "technical efficiency; (2) new ways of determining "returns-to-scale efficiencies; and (3) new approaches for evaluating "allocative efficiencies." A final section outlines areas where additional research is needed.

1. Introduction:

DEA can best be described as data-oriented in that it effects performance evaluations and other inferences directly from observed data with minimal assumptions . A tide of reports on its uses and extensions has developed and this has been increasing, at an increasing rate, since publication of the article by Charnes, Cooper and Rhodes [16] in the *European Journal of Operations Research.* The bibliographies released periodically by L.M. Seiford [28] are now approaching 1,000 references. Included are numerous references to uses of DEA to evaluate the performances of not-for-profit and governmental entities, with DEA being applied to activities that have proved resistant to other methods of inference and evaluation. Examples include the activities of schools and universities, military services, hospitals and court systems and, more recently, evaluations of the performances of whole economic and social systems. See [24] and [25].

The developments in DEA involve OR tools, such as mathematical programming, which have been adapted to uses involving inferences from already generated data (*ex post*) as distinguished from more traditional uses, as in *ex ante* planning. This is accomplished in a manner that can be described as follows: Application is to observed data generated from past behavior. Common to all uses of DEA are the choices of (a) inputs and outputs from which evaluations are to be effected and (b) the choices of DMUs (=Decision Making Units) that represent the organizations which are to be evaluated relative to each other. The result is a new method for effecting inferences which replaces statistical (or other) averages obtained from optimizations over all observations in order to obtain inferences which are

* Graduate School of Business, The University of Texas at Austin, USA
** Graduate School of Policy Science, Saitama University, Japan
*** Paper to be presented at the 14th European Conference on Operational Research to be held at Hebrew University, Jerusalem, Israel, July 3-6, 1995.

optimal for each observation. Opportunities are thereby opened for complementing as well as competing uses with other approaches for effecting evaluations and measuring efficiency. See Bardhan [9].

Different choices of DMUs can lead to different results (and yield different insights) and the same is true of the input and output choices. These choices are important, of course, but they also offer possibilities that can be exploited in a variety of ways which include expanding or contracting the number of inputs and outputs and DMUs as well as other forms of sensitivity analysis. See Thompson, Dharmapala and Thrall [31]. Weights can also be used or suitable aggregates can be assembled from initially designated inputs, outputs and DMUs. *A priori* choices of weights, however, are not required by DEA and, when desired, a use of exact weights may be replaced with upper or lower bounds while allowing DEA to determine a best set of exact values directly from records of past performance. See Cooper, Tone, Takamori and Sueyoshi [18].

2. Measures of Efficiency:[1]

Our discussions here start with a measure of efficiency that has recently been introduced into the literature of DEA by Thompson, Dharmpala, and Thrall [31], viz.,

$$\max_{u,v} \frac{\sum_{r=1}^{s} u_r y_{ro}}{\sum_{i=1}^{m} v_i x_{io}} \Bigg/ \frac{\sum_{r=1}^{s} u_r y_{rk}}{\sum_{i=1}^{m} v_i x_{ik}} .$$

where (1)

$$\frac{\sum_{r=1}^{s} u_r y_{rk}}{\sum_{i=1}^{m} v_i x_{ik}} = \max \left\{ \frac{\sum_{r=1}^{s} u_r y_{rj}}{\sum_{i=1}^{m} v_i x_{ij}} \;\middle|\; j = 1, \ldots, n \right\}$$

which we refer to as the TDT measure of relative efficiency. Here (u,v) are vectors with components $u_r, v_i \geq 0$, with values to be determined by DEA from the observed values of i = 1, ..., m inputs used and r = 1, ..., s outputs produced for each of j = 1, ..., n DMUs. DMU_o, as represented in the numerator for the objective in (1), is the DMU to be evaluated by choosing (u,v) to maximize its value relative to the highest score that this same (u,v) choice accords to the similarly formed ratios for the entire collection of DMU_j, j = 1, ..., n—with DMU_o included in this collection.

[1]The materials in this section are adapted from Banker and Cooper [7].

Formally we can also represent the ratio of ratios in the objective of (1) by

$$\frac{y_o}{x_o} \Big/ \frac{y_k}{x_k} \tag{2}$$

where y_o, y_k represent "virtual outputs" and x_o, x_k represent "virtual inputs." Because y_k/x_k is maximal over the set $k = 1, \ldots, n$, which includes $k = o$, we have $y_o/x_o \leq y_k/x_k$. The above ratios therefore have a maximum value of unity which is achievable if and only if DMU_o's performance is not bettered by some other DMU.

Note, therefore, that the representation in (1) produces a new method of choosing weights. These weights are not fixed *à priori*. Rather a best set of weights is to be determined for each DMU to be evaluated in the sense that the weights selected should maximize the objective in (1) for each DMU that is to be evaluated. This choice of weights is effected directly from the data in a manner which guarantees that the resulting choice is best relative to the best score that these same weights will assign to the other DMUs (including the DMU_o to be evaluated).

This TDT measure may be interpreted in various ways. It might be regarded as an extreme value statistic, for example, and treated by suitably extended versions of extreme value statistical theory. The formulation in (1) may also be approached deterministically as a mathematical programming problem to be modeled and solved in ways suited to choosing the u and v vectors, and this is the way in which we now proceed. The following model, known as the CCR ratio model,[2] can clarify what is involved,

$$\begin{aligned}
&\max_{u,v} \quad \frac{\sum_{r=1}^{s} u_r y_{ro}}{\sum_{i=1}^{m} v_i x_{io}} \\
&\text{subject to} \\
&\frac{\sum_{r=1}^{s} u_r y_{rj}}{\sum_{i=1}^{m} v_i x_{ij}} \leq 1; \qquad j=1, \ldots, n. \\
&\frac{u_r}{\sum_{i=1}^{m} v_i x_{io}} \geq \varepsilon; \qquad r=1, \ldots, s. \\
&\frac{v_i}{\sum_{i=1}^{m} v_i x_{io}} \geq \varepsilon; \qquad i=1, \ldots, m.
\end{aligned} \tag{3}$$

[2]See [16]. See also E. Rhodes [27].

Here the only new element is ε, a positive Non-Archimedean infinitesimal. We elaborate on its mathematical properties later after noting that its use ensures that all inputs and outputs are accorded "some" positive value. These values need not be specified explicitly but can be dealt with by computational processes like the ones described in [2].

As shown in [16], the formulation in (3), greatly generalizes the usual single-output-to-single-input ratio definitions of efficiency that are used in engineering and science. It also relates these engineering-science definitions and usages to definitions in economics—e.g., the Pareto-Koopmans definitions of efficiency given in Charnes *et al.* [13]—which can be accorded operationally implementable form via the following definition,

> Efficiency: The performance of DMU_0 is to be considered fully (100%) efficient if and only if the performance of other DMUs does not provide evidence that some of the inputs or outputs of DMU_0 could have been improved without worsening some of its other inputs or outputs.

We will shortly provide a transformation of (3) that makes it possible to identify the sources and estimate the amounts of inefficiency in each input and output for every DMU in a manner that uses minimal assumptions for empirical studies. Here we note that a relation to (1) is established by simply observing that a necessary condition for optimality in (3) is that at least one of the $j = 1, ..., n$ output-to-input ratios in the constraints must be at its upper bound of unity. Thus, for this case the denominator in (1) has a value of unity and the efficiency evaluation for DMU_0 then reduces to whether the numerator in (1) or (2) is unity or less.

Maximizing y_0/x_0 can be managerially interpreted in terms of achieving the greatest virtual output per unit virtual input. This provides a basis for extending DEA to consider returns to scale efficiencies, as we shall later see. Here, however, we simply note that this interpretation corresponds mathematically to finding values which can be associated with the slopes of the supports that envelop the observations. Nothing need be said explicitly about the functions that govern the relations between inputs and outputs and these relations are allowed to vary from one DMU to another.

3. Linear Programming Equivalents:

Reference to (3) shows that it is a nonlinear, nonconvex programming problem, and hence is best used for conceptual clarification. To give these concepts computationally implementable form, we introduce new variables defined as follows,

$$\mu_r = t\, u_r, \quad r = 1, \ldots, s$$
$$\nu_i = t\, v_i, \quad i = 1, \ldots, m \tag{4}$$
$$1 = \sum_{i=1}^{m} \nu_i\, x_{io}.$$

These are the so-called "Charnes-Cooper transformations" from Charnes and Cooper [12] which initiated the field of "fractional programming." Here we use them to transform the problem in (3) to the problem on the right in the following dual pair of linear programming problems with assurance (from fractional programming) that their optimal values will also be optimal for (3).

$$\begin{array}{ll}
\min\ \theta \quad -\varepsilon\left(\sum_{i=1}^{m} s_i^- + \sum_{r=1}^{s} s_r^+\right) & \max \quad \sum_{r=1}^{s} \mu_r\, y_{ro} \\
\text{subject to:} & \text{subject to:} \\
0 = \theta x_{io} - \sum_{j=1}^{n} x_{ij}\lambda_j - s_i^- & -\sum_{i=1}^{m} \nu_i\, x_{ij} + \sum_{r=1}^{s} \mu_r\, y_{rj} \le 0 \\
y_{ro} = \sum_{j=1}^{n} y_{rj}\lambda_j \quad - s_r^+ & \sum_{i=1}^{m} \nu_i\, x_{io} = 1 \\
0 \le \lambda_j,\ s_i^-,\ s_r^+, & -\nu_i \le -\varepsilon \\
 & -\mu_r \le -\varepsilon
\end{array} \tag{5}$$

where $i = 1, \ldots, m$ indexes the inputs, $r = 1, \ldots, s$ indexes the outputs, and $j = 1, \ldots, n$ indexes the DMUs. As before, $j = o$ is used to identify the DMU to be evaluated by (a) placing its data in the objective while also (b) leaving it in the constraints. Leaving the data for DMUo in the constraints guarantees that solutions exist for both problems in (5) and, by the dual theorem of linear programming, it follows that they will have finite and equal optimal values.

We now assume that the x_{ij}, y_{rj} are all positive[3] so that we can move back and forth between (5) and (3) because the constraint $\sum_{i=1}^{m} \nu_i\, x_{io} = 1$, $\nu_i \ge 0$, all i, means that we have $t > 0$ in (5). Hence we have the full power of available linear programming algorithms and computer codes to solve (5) or (3), as we wish. We also have its interpretative power available (after suitable adaptations) for use in DEA efficiency analyses and inferences.

Using * to denote an optimal value, the condition for full (100%) DEA efficiency as defined above becomes:

$$\sum_{r=1}^{s} \mu_r^*\, y_{ro} = 1 \tag{6}$$

---

[3]This condition can be relaxed. See A. Charnes, W. W. Cooper and R. M. Thrall [17].

for the problem on the right in (5). Interest usually attaches to identifying sources and amounts of inefficiency in each input and output. This is most easily done from the problem on the left where the conditions for efficiency become,

$$\begin{aligned} &\text{(i) } \theta^* = 1 \\ &\text{(ii) All optimum slack values are zero.} \end{aligned} \tag{7}$$

It is to be noted that the presence of non-zero slacks means that the measure of inefficiency resulting from (7) involves a two-component number of the form $\theta^* - k^* \varepsilon$ where k^* = sum of slacks. Both θ^* and k^* are real numbers and hence are Archimedean, whereas ε is a Non-Archimedean infinitesimal so that $\theta^* - k^* \varepsilon$ is not a real number unless $k^* = 0$.

One way to achieve a single real-number measure of inefficiency is to confine attention to "weak efficiency" and thereby ignore the slacks. Also referred to as "Farrell" or "Farrell-Debreu" efficiency—see references—its use corresponds to an assumption of "free disposal" so that, mathematically, $\varepsilon > 0$ is replaced by zero and the property of maximizing the slacks after ascertaining an optimal θ^* is thereby lost. See Färe, Grosskopf and Lovell [21]. However, when slacks are an important source of possible inefficiency (e.g., because inefficient mixes have been used) then alternate approaches may need to be considered. The measures referred to as MID and MED in Bardhan *et al.* [10] might then be adapted to more general situations by using what are referred to as the CCR projections which we describe in the following manner.

First observe that the Non-Archimedean element $\varepsilon > 0$ is not present in the constraints for the problem on the left in (5). Hence the values in the constraints involve only real numbers. Thus, when an optimal solution is available we have,

$$\theta^* x_{io} - s_i^{-*} = \sum_{j=1}^{n} x_{ij} \lambda_j^* = x_{io}^*$$

where x_{io}^*, $i = 1, \ldots, m$, represents the value of the i^{th} virtual input for DMU_o. It therefore follows that

$$x_{io} - x_{io}^* \leq x_{io} \quad \text{or} \quad \frac{x_{io} - x_{io}^*}{x_{io}} \leq 1, \tag{8}$$

and, similarly,

$$y_{ro} + s_r^{+*} = y_{ro}^* \quad \text{or} \quad \frac{y_{ro}^* - y_{ro}}{y_{ro}^*} \leq 1 \tag{9}$$

from which we can derive the following measure of inefficiency,

$$0 \le \frac{\sum_{i=1}^{m} \frac{x_{io} - x_{io}^*}{x_{io}} + \sum_{r=1}^{s} \frac{y_{ro}^* - y_{ro}}{y_{ro}^*}}{s + m} \le 1 \tag{10}$$

where s is the number of outputs and m is the number of inputs. As can be seen, this represents an average of the inefficiency proportions due to (i) excessive inputs in the first term and (ii) output shortfalls in the second term. To convert this to a measure of efficiency we replace the above with

$$0 \le 1 - \frac{\sum_{i=1}^{m} \frac{x_{io} - x_{io}^*}{x_{io}} + \sum_{r=1}^{s} \frac{y_{ro}^* - y_{ro}}{y_{ro}^*}}{s + m} \le 1. \tag{11}$$

Referring to the measure in (10) as MIP (Measure of Inefficiency Proportions) and the measure in (11) as MEP (Measure of Efficiency Proportions) we see that full efficiency is attained only with MEP = 1 or MIP = 0.

This now provides a way of reducing the conditions represented in (6) or (7) to a single real number without losing the ability to identify the inefficiencies that may be present in inputs, in outputs, or in subsets thereof.[4] Extensions to weighted measures and the treatment of zeros which may appear to be sources of trouble in some of the denominators are discussed in Bardhan *et al.* [10]. We therefore do not cover these topics here because this treatment requires only very natural extensions of the MID and MED measures discussed in Bardhan *et al.* [11] as well as in Banker and Cooper [7]. We do need to note, however, that these measures also lend themselves to rankings of DMU performances whereas this is not the case for the often used θ^* values obtained from (5) because (a) the latter measure is incomplete and (b) these θ^* values may be determined from different facets—which means that these values are being derived from comparisons involving performances of different sets of DMUs.

4. Returns to Scale:

We have just completed our discussion of some of the more recent developments like the TDT and MEP measures of "technical efficiency" in DEA. Next we turn to some of the recent developments for uses of DEA to identify "returns-to-scale" situations and we initiate this discussion with the following theorem from Banker and Thrall [8] which is to be used with the primal model in (5):

[4]An analogous measure has been developed by C.A.K. Lovell and J. Pastor [25] which they refer to as GEM (=Generalized Efficiency Measure).

CCR Returns to Scale Theorem (Banker and Thrall):

If $\sum_{j=1}^{n} \lambda_j^* = 1$ in *any* alternate optima then constant returns to scale prevail.

If $\sum_{j=1}^{n} \lambda_j^* > 1$ for *all* alternate optima then decreasing returns to scale prevail.

If $\sum_{j=1}^{n} \lambda_j^* < 1$ for *all* alternate optima then increasing returns to scale prevail.

where "*" refers to an optimal value for the DMU_o being evaluated.

Banker and Thrall assume that the input-output vector (X_o, Y_o) for the DMU_o to be evaluated is positioned on the frontier which is "radially (i.e., Farrell-Debreu) efficient. More recent papers by Banker, Chang and Cooper [4] and [5] show how this assumption may be eliminated.

We illustrate this BCC procedure by proceeding as follows: Suppose we have an optimum solution to (5) with $\sum_{j=1}^{n} \lambda_j^* < 1$. To check on alternate optima possibilities we utilize

$$\max \sum_{j=1}^{n} \hat{\lambda}_j + \varepsilon\left(\sum_{i=1}^{m} \hat{s}_i^- + \sum_{r=1}^{s} \hat{s}_r^+\right)$$

subject to

$$\begin{aligned} \theta_o^* X_o &= \sum_{j=1}^{n} X_j \hat{\lambda}_j + \hat{s}^- \\ Y_o &= \sum_{j=1}^{n} Y_j \hat{\lambda}_j - \hat{s}^+ \\ 1 &\geq \sum_{j=1}^{n} \hat{\lambda}_j , \end{aligned} \tag{12}$$

where θ_o^* is the optimal value for (5) and the slack vectors $\hat{s}^-$ and $\hat{s}^+$, as well as the components of the vector $\hat{\lambda}$, are constrained to be non-negative.

As we shall shortly see, we can dispense with the non-Archimedean terms in the objective for (5). Because it is incorporated in (12) we can obtain an optimal solution from (5) without worrying about slack inefficiencies and then utilize (12) for which we have the following theorem from [4]:

Theorem: Given the existence of an optimal solution with $\sum_{j=1}^{n} \lambda_j^* < 1$ in (5), the returns to scale associated with (X_o, Y_o) will be constant if and only if $\sum_{j=1}^{n} \hat{\lambda}_j^* = 1$ and increasing if and only if $\sum_{j=1}^{n} \hat{\lambda}_j^* < 1$ in (12).

We are here restricting attention to solutions of (12) with $\sum_{j=1}^{n} \hat{\lambda}_j \leq 1$ because we are considering the case of an optimal solution to (5) with $\sum_{j=1}^{n} \lambda_j^* < 1$. To consider the alternative situation of $\sum_{j=1}^{n} \lambda_j^* > 1$ we simply reorient the objective and reverse the inequality to $\sum_{j=1}^{n} \hat{\lambda}_j \geq 1$ in (12). Finally, nothing further is required if $\sum_{j=1}^{n} \lambda_j^* = 1$ in (5).

For illustration, we use Figure 1, below, to which we assign the following coordinate values,

$$A = (1, 1),\ B = (\tfrac{3}{2}, 2),\ C = (3, 4),\ D = (4, 5),\ E = (4, 4\tfrac{1}{2}) \tag{13}$$

with the first component an input and the second an output. The points A, B, C and D are all on the solid line that represents the efficiency frontier for the "BCC model"—see Banker, Charnes and Cooper [6]—which is the same as primal model on the left in (5) with the convexity condition $\sum_{j=1}^{n} \hat{\lambda}_j = 1$ adjoined. The ray represented by the broken line from the origin represents the efficiency frontier for the CCR model.

The region from B to C lies in the intersection of the two frontiers, and <u>this is not true for any other efficient point</u>. This is referred to as the region of MPSS (=Most Productive Scale Size) because it is where the output to input ratio is maximal. This follows readily from

$$\frac{d(y/x)}{dx} = \frac{x\frac{dy}{dx} - y}{x^2} = 0$$

so that for $x \neq 0$ we have marginal product equal to average product

$$\frac{dy}{dx} = \frac{y}{x}.$$

This, in turn, gives

$$\frac{x}{y}\frac{dy}{dx} = \frac{d \ln y}{d \ln x} = 1$$

so that proportional increases in input are associated with equi-proportional increases in output and returns to scale are constant in this region. Finally we now note that a point like B or C, which is CCR efficient will also be BCC efficient, but the converse is not necessarily true. See [1]

We illustrate by using the coordinates of A in (13). Substituting in (5) from the data in (13) gives

$$\min \theta_o$$

subject to

$$1\theta_o \geq 1\lambda_A + \tfrac{3}{2}\lambda_B + 3\lambda_C + 4\lambda_D + 4\lambda_E$$
$$1 \leq 1\lambda_A + 2\lambda_B + 4\lambda_C + 5\lambda_D + \tfrac{9}{2}\lambda_E \qquad (14)$$
$$0 \leq \lambda_A, \lambda_B, \lambda_C, {}_D, \lambda_E.$$

This problem has min $\theta_o = \theta_o^* = \frac{3}{4}$ with alternate optima represented by $\lambda_B^* = \frac{1}{2}$ or $\lambda_C^* = \frac{1}{4}$ and all other $\lambda^* = 0$. In either case, we have $\sum_{j=1}^{n} \lambda_j^* < 1$, so we utilize (12) to write

$$\max \hat{\lambda}_A + \hat{\lambda}_B + \hat{\lambda}_C + \hat{\lambda}_D + \hat{\lambda}_E + \varepsilon(\hat{s}^- + \hat{s}^+)$$

subject to

$$\tfrac{3}{4} = 1\hat{\lambda}_A + \tfrac{3}{2}\hat{\lambda}_B + 3\hat{\lambda}_C + 4\hat{\lambda}_D + 4\hat{\lambda}_E + \hat{s}^- \qquad (15)$$
$$1 = 1\hat{\lambda}_A + 2\hat{\lambda}_B + 4\hat{\lambda}_C + 5\hat{\lambda}_D + \tfrac{9}{2}\hat{\lambda}_E - \hat{s}^+$$
$$1 \geq \hat{\lambda}_A + \hat{\lambda}_B + \hat{\lambda}_C + \hat{\lambda}_D + \hat{\lambda}_E .$$

The solution with this objective gives $\hat{\lambda}_B^* = \frac{1}{2}$ and all other variables zero. It follows from the above theorem that increasing returns to scale prevails at A in Figure 1.

As previously observed, assigning the non-Archimedean element to the slacks in (15), makes it unnecessary to use it in (5). We also note that we have eliminated the need for examining all alternate optima, as required by the Banker-Thrall theorem, since our solution with $\sum_{j=1}^{n} \hat{\lambda}_j^* < 1$ in (15) suffices to establish that no solution is available with $\sum_{j=1}^{n} \hat{\lambda}_j \geq 1$ and $\theta_o^* = 3/4$.

Although A is not MPSS, such a point is readily obtained from the following projection operators due to Banker and Morey [8],

$$\hat{x}_{io}^* = \frac{\theta_o^* \, x_{io} - \hat{s}_i^{-*}}{\sum_{j=1}^{n} \hat{\lambda}_j^*}, \qquad i = 1, \ldots, m \qquad (16)$$

$$\hat{y}_{ro}^* = \frac{y_{ro} + \hat{s}_r^{+*}}{\sum_{j=1}^{n} \hat{\lambda}_j^*}, \qquad r = 1, \ldots, s,$$

where the $\sum_{j=1}^{n} \hat{\lambda}_j^*$ and the $\hat{s}_i^{-*}$ and $\hat{s}_r^{+*}$ refer to optimal solutions for (12) and θ_o^* is the value transferred from (5) to (12). Thus, using our solution to (15) we obtain

$$\hat{x}_{io}^* = \frac{\frac{3}{4} - \hat{s}^{-*}}{\frac{1}{2}} = \frac{3}{2}$$

$$\hat{y}_{ro}^* = \frac{1 + \hat{s}^{+*}}{\frac{1}{2}} = 2 \tag{17}$$

and observe from the coordinates given in (13) that this projects into B in Figure 1 which is MPSS. Although it is not necessary to do so, we use the alternate optimum with $\lambda_C^* = \frac{1}{4}$ for $\sum_{j=1}^{n} \hat{\lambda}_j^*$ and obtain

$$\hat{x}_o^* = \frac{3}{4} \Big/ \frac{1}{4} = 3$$

$$\hat{y}_o^* = 1 \Big/ \frac{1}{4} = 4 \tag{18}$$

which projects into C—also MPSS—and, of course, convex combinations of these alternate optima may be used to obtain other points between B and C which are <u>all</u> MPSS.

Now we turn to E in (7) which is <u>not</u> on either the CCR or the BCC efficiency frontiers. To evaluate E, however, we simply alter the terms on the left in (15) to obtain

$$\begin{aligned}
&\min \theta_o \\
&\text{subject to} \\
&\quad 4\theta_o \geq 1\lambda_A + \tfrac{3}{2}\lambda_B + 3\lambda_C + 4\lambda_D + 4\lambda_E \\
&\quad \tfrac{9}{2} \leq 1\lambda_A + 2\lambda_B + 4\lambda_C + 5\lambda_D + \tfrac{9}{2}\lambda_E \\
&\quad 0 \leq \lambda_A, \lambda_B, \lambda_C, \lambda_D, \lambda_E.
\end{aligned} \tag{19}$$

Again we have alternate optima with, now, $\theta_o^* = \frac{27}{32}$ for either $\lambda_B^* = \frac{9}{4}$ or $\lambda_C^* = \frac{9}{8}$ and all other $\lambda^* = 0$. Hence, $\sum_{j=1}^{n} \lambda_j^* > 1$ in both cases. So, proceeding to the indicated modification of (12), we form

$$\begin{aligned}
&\min (\hat{\lambda}_A + \hat{\lambda}_B + \hat{\lambda}_C + \hat{\lambda}_D + \hat{\lambda}_E) - \varepsilon(\hat{s}^- + \hat{s}^+) \\
&\text{subject to} \\
&\quad \tfrac{27}{8} = 1\hat{\lambda}_A + \tfrac{3}{2}\hat{\lambda}_B + 3\hat{\lambda}_C + 4\hat{\lambda}_D + 4\hat{\lambda}_E + \hat{s}^- \\
&\quad \tfrac{9}{2} = 1\hat{\lambda}_A + 2\hat{\lambda}_B + 4\hat{\lambda}_C + 5\hat{\lambda}_D + \tfrac{9}{2}\hat{\lambda}_E - \hat{s}^+ \\
&\quad 1 \leq \hat{\lambda}_A + \hat{\lambda}_B + \hat{\lambda}_C + \hat{\lambda}_D + \hat{\lambda}_E \\
&\quad 0 \leq \hat{\lambda}_A, \hat{\lambda}_B, \hat{\lambda}_C, \hat{\lambda}_D, \hat{\lambda}_E.
\end{aligned} \tag{20}$$

This has its optimum at $\hat{\lambda}_C^* = \frac{9}{8}$ with all other variables equal to zero. We therefore conclude that decreasing returns prevail for this point or, more precisely, at the point on the frontier to which this solution is projected by (8) and (9).

Again, for further insight, we use (16) to obtain

$$\hat{x}_0^* = \frac{27}{8} \Big/ \frac{9}{8} = 3, \qquad \hat{y}_0^* = \frac{9}{2} \Big/ \frac{9}{8} = 4, \tag{21}$$

and reference to (13) shows that these are the coordinates for C in Figure 1. Similarly an application to $\lambda_B^* = \frac{9}{4}$, the alternate optimum to (20), gives

$$\hat{x}_0^* = \frac{27}{8} \Big/ \frac{9}{4} = \frac{3}{2}, \qquad \hat{y}_0^* = \frac{9}{2} \Big/ \frac{9}{4} = 2 \tag{22}$$

which are the coordinates for B.

We now note that C and B are also basis members with positive coefficients in all of our optima where they serve as the reference DMUs in the set of efficient DMUs used to evaluate other DMUs. These are points in the interval that is common to both the BCC and CCR efficiency frontiers. These points and their convex combinations are MPSS, and these are the only points that are both CCR and BCC efficient.

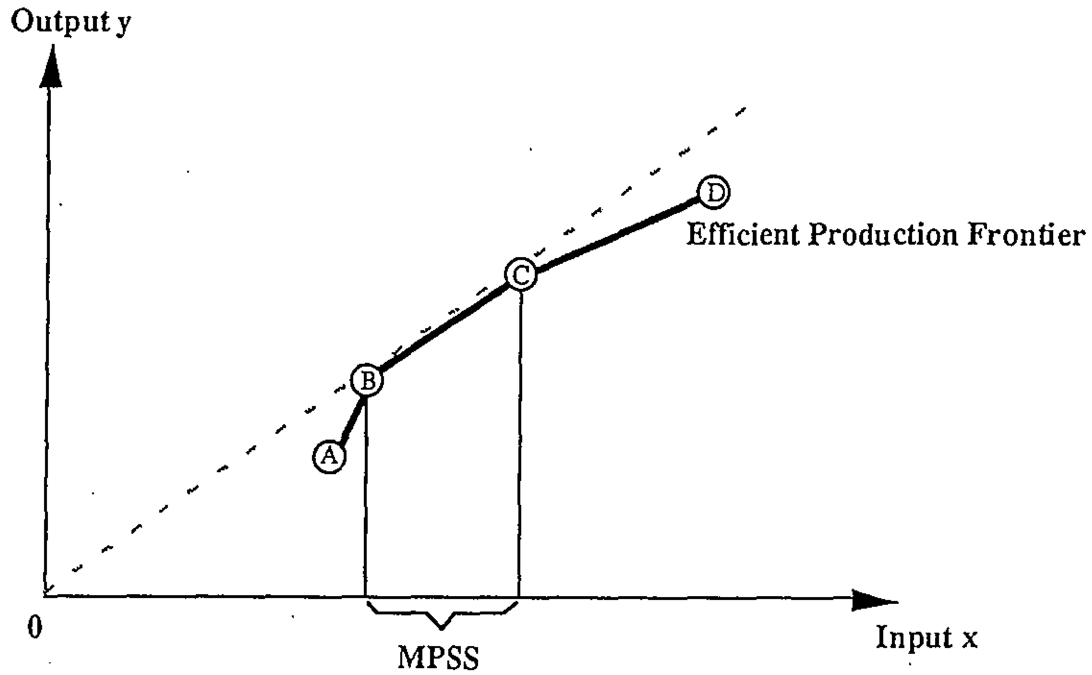

Figure 1: MPSS = Most Productive Scale Size

Using the above examples for guidance we can see that the situation is general from the following considerations. First, as already remarked, a point which is found to be efficient for the CCR model will also be efficient for the BCC model (but the converse is not necessarily true). Second, as shown in [3] a DMU can be part of an optimal basis with a positive coefficient only if it is efficient. Hence all members of an optimal basis with positive coefficients lie in the intersection of the CCR and BCC efficiency frontiers. The rest of what is needed follows from the other properties of the efficiency frontiers of both the BCC and CCR models--viz., such frontiers are concave, monotonically increasing and continuous. In short, the associated frontier function is "isotonic" in the manner described in Charnes et *al.* [13].

The following mathematical formulation shows what is happening analytically. When (5) and (12) are used to evaluate DMU_0 with input vector X_0 and output vector Y_0 we can write the optimal solution as

$$P_E = \begin{pmatrix} \theta_0^* X_0 - \hat{s}^{+*} \\ Y_0 + \hat{s}^{-*} \end{pmatrix} = \sum_{j=1}^{n} P_j \hat{\lambda}_j^* . \qquad (23)$$

Here P_E is a point on the efficient production frontier. It is the point used to evaluate DMU_0 with coordinate values the same as those obtained by using the CCR projection operators defined in (8) and (9). In short, the $\hat{\lambda}_j^*$ in (23) effect this same projection and so our returns-to-scale characterizations are also obtained from these $\hat{\lambda}_j^*$ values for points on the efficient production frontier. Because P_E is on the CCR efficiency frontier it has constant returns-to-scale. Hence the Banker-Morey projection in (16) may be used to bring P_E into MPSS, when desired, with no need to adjust the observed mix to the efficiency proportions required at MPSS because this has already been accomplished by (23). In any case, the value of $\sum_{j=1}^{n} \hat{\lambda}_j^*$ serves to relate the scale at which P_E operates to what is required for MPSS with $\sum_{j=1}^{n} \hat{\lambda}_j^* < 1$ when expansion is indicated and $\sum_{j=1}^{n} \hat{\lambda}_j^* > 1$ when contraction is needed to achieve MPSS.

Using Figure 1 to review what has been said we observe that E was first projected onto the CCR frontier which is both technically and scale efficient. Points in the region of MPSS are the only ones that can enter an optimal basis with positive coefficients. Hence the sum of these coefficients indicates whether the thus projected point is above, below or in this region. Evaluation of the returns to scale for the corresponding BCC model can thus be effected from the properties of concavity and continuity associated with the efficiency frontier for the latter models.

We should perhaps note that there may be more than one P_E if alternate optimum are present in (12). This presents no problem because the same optimal value for the sum of the $\hat{\lambda}$s will be applicable. However, as noted in Banker and Thrall [9], the situation for returns to scale may change if an "output-oriented" version of the CCR model is used. In important applications, it may therefore be a good idea to use both the "input" and the "output oriented" versions of these models and choose between them according to whether output augmentation or input conservation is to be emphasized.

Here we have focused on the CCR model which evaluates both technical and returns-to-scale efficiency. The BCC model separates the two—using the primal model to evaluate technical efficiency and its dual to evaluate returns-to-scale efficiency. Either the CCR or BCC model may be used to determine the situation for returns to scale locally. However, both must be used if quantitative estimates are to be made of benefits that are obtainable when increasing or decreasing returns are present.

5. Allocatively Most Efficient Points and Regions:[5]

When moving from section 3 to section 4, we turned from "technical inefficiencies," to "returns-to-scale inefficiencies." The latter can involve movement along the efficiency frontier to achieve, say, MPSS. This is not the case for technical inefficiencies because in this case movement is to the frontier rather than along it. Technical inefficiency, when present, may thus be regarded as "waste" because no substitutions are required for their removal. Movement from a point on the efficiency frontier to achieve MPSS will require input-output exchanges with corresponding "tradeoffs."

We now turn to "allocative inefficiency," which represents a third member of the three categories introduced by M.J. Farrell [22] and [23]. This involves movement along the efficiency frontier with tradeoffs occurring in response to criteria like unit prices, unit costs, unit profits or utility theoretic weight assignments. Unfortunately requisite information in the form of exact values of "prices," etc., are often not available in many applications. However, research in DEA has begun to provide alternatives in which inequalities may be used to place upper and lower bounds on ranges rather than the exact values that would be required to determine "Allocatively Most Efficient" (AME) points. One

[5]The material in this section has been adapted from Cooper, Tone, Takamori and Sueyoshi [18].

example of results from such research is represented by the "assurance regions" which Thompson, Singleton, Thrall and Smith [30] introduced to help locate a site for the "Superconducting Super Collider." See also Dyson and Thanassoulis [20]. Another example are "cone ratio envelopments" as presented in Charnes *et al.* [14] and [15] to monitor the performances of individual banks in a regulatory system.

To describe these "cone ratio envelopments we use Figure 2, in which the models in (5) have been reoriented to "output maximization" rather than "input minimization"—and we schematically portray the outputs of 5 DMUs by associating them with the points P_1, P_2, P_3, P_4 and P_5 where the amounts of each of their two outputs are represented by the coordinate values (y_1, y_2).

Using (w_1, w_2) to represent fixed values such as unit profits, or some similar measure of desirability, we can form the expressions

$$w_1 y_1 + w_2 y_2 = k,$$

or

$$y_2 = \frac{k}{w_2} - \frac{w_1}{w_2} y_1 \tag{24}$$

in order to rank the choices between y_1 and y_2 according to the value of k. Here we can simplify matters by assuming that the costs incurred and the prices received for each of the.two products are the same for every DMU so that revenues (and hence profits) vary at the same w_1, w_2 rate for each of the output choices.

The broken lines which slant through P_3 and P_1 in Figure 2 represent two possibilities for the (y_1, y_2) choices that are associated with these lines. Because k_2 exceeds k_1, movement from P_3 to P_1 is justified on these (w_1, w_2) criteria, even though P_3 is technically efficient.

We are here examining the property of "allocative efficiency" and observe that when relative values are specified, as in the choice of (w_1, w_2), it is possible to determine efficient points which are better than others even when the latter are also efficient. Indeed, a parallel movement of the broken line through P_3 to a position where it goes through P_4 shows the latter to be preferred even though it is not on the efficiency frontier.

Evidently the "technical efficiency" achieved by P_3 is not sufficient to guarantee allocative efficiency. However, being on the efficiency frontier is a necessary condition for allocative efficiency. If a point is not on the efficiency frontier, there will always be some point which is allocatively more efficient. In the situation of Figure 2, in fact, P_1 exhibits the best possible resource allocation when the criteria of relative choice are given by the values (w_1, w_2) associated with the broken lines in Figure 2.

We next explore possible variations in the values of (w_1, w_2) that leave P_1 invariant as the most preferred—i.e., allocatively most efficient choice. The arrows represent "normals" to the line segments shown in Figure 2, and these can be related to the values of (w_1, w_2) in the following way. The dotted arrows through P_1 and P_3 are obtained from the values of (w_1, w_2) which yield a direction that is perpendicular to the broken lines labeled $\frac{k_2}{w_2}$ and $\frac{k_1}{w_2}$, respectively. Hence, the arrows represent the <u>direction</u> in which optimization is to be undertaken. No further movement in this direction is possible at P_1 without leaving the set of production possibilities. Hence P_1 is optimal, whereas this is not the case for P_3 since there are other (y_1, y_2) choices available on this line which are well inside the set of production possibilities. The solid arrows in Figure 2 represent normals to the solid line segments with which they are associated. These normals represent the rates at which substitutions must be made between y_1 and y_2 in order to stay on that segment of the frontier.

The two solid arrows on either side of the dotted arrow at P_1 set limits to its allowed variations. Changing the tilt of the broken line will not change P_1 from its status as allocatively efficient so long as the normal associated with each such tilt does not go outside the solid arrows which surround it. Only when the tilt is sufficient to produce a dotted arrow that breaches one of these limits is movement to a new, allocatively most efficient, point justified. This new allocatively most efficient point can then be found from the new values (w_1', w_2') which indicate the new direction of optimization associated with this breach.

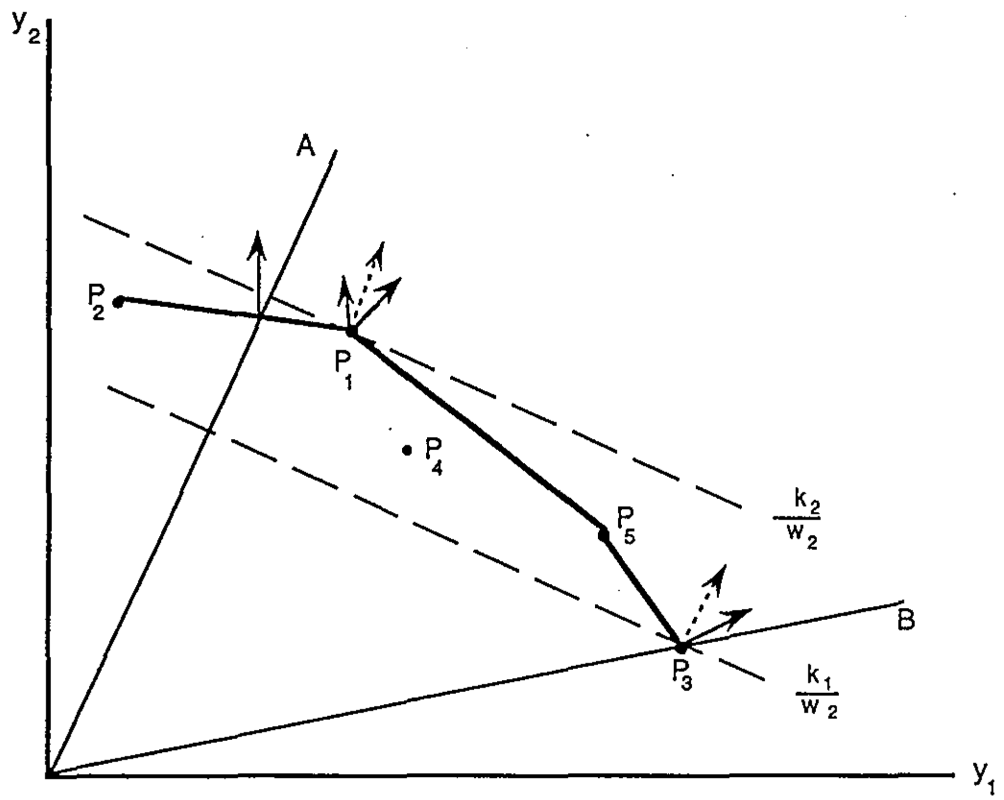

FIGURE 2
ALLOCATIVE EFFICIENCY

Thus far we have proceeded on the assumption of known values for (w_1, w_2) which form the components of the vector that is normal to (24). On this assumption, it becomes possible to identify an AME (=allocatively most efficient) point. However, the assurance region and cone ratio envelopments make it possible to deal with cases in which the (w_1, w_2) values are not known precisely. This is accomplished by replacing the search for an AME point by an Allocatively Most Preferred (AMP) region which lies within limits prescribed for allowable choices of the dual variables.

Turning once more to Figure 2, the solid lines represented by A and B form the edges of a cone that delimit the region of choice. It is only the points inside this region that lie in the production possibility set. In the cone-ratio envelopment approach, the points to be chosen must lie on segments of the efficiency frontier with normals that point in directions for movement that remain inside the cone. The normal on the segment between P_1 and P_2 points outside the cone, as shown at the intersection with line A. Hence the segment of the efficient frontier between P_2 and P_1 is eliminated from consideration. However, P_1 remains a candidate for designation as an efficient performer as do P_5 and P_3—and points on any of the line segments that form these parts of the efficient frontier.

Tighter limits may be imposed with, in the limit, allocative efficiency being achieved, if possible. Alternatively iterative approaches may be used as is done in Charnes, Cooper, Huang and Sun [14] and [15] where a set of (supposedly) excellent banks are designated for use in forming the cones by reference to their (optimal) dual variable values. This set of excellent banks is then tested for efficiency and the ones that fail this test are eliminated from consideration, and so on.

These approaches also lend themselves to new types of sensitivity analyses as shown by Thompson, Dharmapala and Thrall [31] who use their assurance region concepts to introduce a new method of sensitivity analysis that allows all data to vary until a change from efficient to inefficient status (or *vice versa*) first occurs for any DMU. A reassignment is then made, after which the analysis is repeated, and so on. In exchange for some fusing of the concepts of technical and allocative efficiency, to attain AMP regions in place of AME points, we see not only a range of potential new applications for DEA but we also see a range of new topics for research being opened.

6. Conclusion:

We have now covered some of the recent developments in DEA that deal with technical, scale and allocative efficiency. In doing so, we have necessarily short changed or hidden from view many other developments. New applications and new analytical developments are now occurring in such a volume and at such a rate, however, as to make it almost impossible to keep abreast of them. We therefore close, instead, by simply listing topics like the following which are urgently in need of further attention: (1) economies of scope (2) efficiency and flexibility tradeoffs (3) efficiency and effectiveness interactions and (4) new non-parametric methods for dealing with uncertain data and/or uncertain choices of inputs, outputs and DMUs.

REFERENCES

[1] Ahn, T., A. Charnes and W.W. Cooper: A Note on the Efficiency Characterizations Obtained in Different DEA Models. *Socio-Economic Planning Sciences* 22 (1988) 253-257.

[2] Arnold, V., I. Bardhan, W.W. Cooper and A. Gallegos: Primal and Dual Optimality in Computer Codes Using Two-Stage Solution Procedures in DEA. In: Jay Aranson and S. Zionts, Eds., *Operations Research:Methods, Models and Applications, A Conference in Honor of G.L. Thompson on His 70th Birthday*. Kluwer Academic Press, Forthcoming, 1995.

[3] Banker, R.D., I. Bardhan, W.W. Cooper and R.M. Thrall: Projection Operators in DEA. *The Annals of Operations Research.* Submitted (1995)

[4] Banker, R.D., H. Chang and W.W. Cooper: Returns to Scale Equivalencies and Their Implementation in DEA. *European Journal of Operational Research* forthcoming, 1995.

[5] Banker, R.D., H. Chang and W.W. Cooper: Simulation Studies of Efficiency, Returns to Scale and Misspecification with Nonlinear Functions in DEA. *The Annals of Operations Research.* To appear (1995).

[6] Banker, R.D., A. Charnes and W.W. Cooper: Models for Estimating Technical and Scale Efficienices in DEA. *Management Science* 30 (1985) 1078-1092.

[7] Banker, R.D. and W.W. Cooper: Validation and Generalization of DEA and Its Uses. *TOP* Sociedad Española de Estadistica e Investigación Operativa, Madrid, 1994.

[8] Banker, R.D. and R.C. Morey: Efficiency Analysis for Exogenously Fixed Inputs and Outputs. *Operations Research* 34 (1986) 513-521.

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

[10] Bardhan, I.: DEA and Stochastic Frontier Regression Approaches to the Evaluation of Public School Performances in Texas. Ph.D. Thesis. The University of Texas at Austin, Graduate School of Business, 1995.

[11] Bardhan, I., W.F. Bowlin, W.W. Cooper and T. Sueyoshi: Models and Measures for Efficiency Dominance and Related Topics in DEA. *Journal of the Operations Research Society of Japan.* Submitted, 1995.

[12] Charnes, A. and W.W. Cooper: Programming with Linear Fractional Functionals. *Naval Research Logistics Quarterly* 9 (1962) 181-186.

[13] Charnes, A., W.W. Cooper, B. Golany, L.M. Seiford and J. Stutz: Foundations of Data Envelopment Analysis and Pareto-Koopmans Efficient Empirical Production Functions. *Journal of Econometrics* 30 (1985) 91-107.

[14] Charnes, A., W.W. Cooper, Z.M. Huang and D.B. Sun: Polyhedral Cone-Ratio DEA Models and Managerial Performance of Large Commercial Banks. *Journal of Econometrics* 46 (1990) 73-91

[15] Charnes, A., W.W. Cooper, Z.M. Huang and D.B. Sun: DEA Approaches to Designs of Bank Monitoring Systems. *Operations Research* submitted, 1992.

[16] Charnes, A., W.W. Cooper and E. Rhodes: Measuring Efficiency of Decision Making Units. *European Journal of Operational Research* 1 (1978) 429-444.

[17] Charnes, A., W.W. Cooper and R.M. Thrall: A Structure for Classifying and Characterizing Efficiency and Inefficiency in Data Envelopment Analysis. *Journal of Productivity Analysis* 2 (1991) 197-237.

[18] Cooper, W.W., K. Tone, H. Takamori and T. Sueyoshi: Survey and Interpretations. *Communications of the Operations Research Society of Japan* (1994) In Japanese. English translations can be secured from the authors.

[19] Debreu, G.: The Coefficient of Resource Utilization. *Econometrica* 19 (1951) 273-292.

[20] Dyson, R.G. and E. Thanassoulis: Reducing Weight Flexibility in DEA. University of Warwick, Warwick Papers in Management. Warwick, England (1987).

[21] Färe, R., S. Grosskopf and C.A.K. Lovell: *The Measurement of Productive Efficiency*. Kluwer-Nijhoff (1985).

[22] Farrell, M.J.: The Measurement of Productive Efficiency. *Journal of the Royal Statistical Society* Series A, 120, Part 3 (1957) 253-280.

[23] Farrell, M.J. and M. Fieldhouse: Estimating Production Functions Under Increasing Returns to Scale. *Journal of the Royal Statistical Society*, Series A (1962) 150-160.

[24] Golany, B. and S. Thore: The Competitiveness of Nations. In W.W. Cooper, D. Gibson, F. Phillips, S. Thore and A. Whinston, eds. IMPACT: *Essays in Honor of George Kozmetsky*. Kluwer Academic Press, 1995, forthcoming.

[25] Lovell, C.A.K. and J.T. Pastor: Macroeconomic Performance of Sixteen Ibero-American Countries over the Period 1980-1981. Working Paper. University of Georgia, Department of Economics, Athens, GA (1994).

[26] Maindiratta, Ajay: Largest Size-Efficient Scale and Size Efficiencies of Decision Making Units in DEA. *Journal of Econometrics* 46 (1990) 57-72.

[27] Rhodes, E.: Data Envelopment Analysis and Related Approaches for Evaluating the Efficiency of Decision Making Units with an Application to Program Follow Through in U.S. Public School Education. Ph.D. Thesis. Carnegie-Mellon University, School of Urban and Public Affairs (1978). Also available from University Microfilms, Inc., Ann Arbor, MI.

[28] Seiford, L.M.: A Bibliography of Data Envelopment Analysis. Technical Report, Department of Industrial Engineering and Operations Research, University of Massachusetts (1994).

[29] Sherman, D.: Measurement of Hospital Technical Efficiency: A Comparative Evaluation of DEA and Other Measurement Techniques for Measuring and Locating Inefficiency in Health Care Organizations. DBA Thesis, Harvard University, Graduate School of Business (1981). Also available from University Microfilms, Inc., Ann Arbor, MI.

[30] Thomas, D.A.: Data Envelopment Analysis Methods in the Management of Personnel Recruitment Under Competition in the Context of U.S. Army Recruiting. Ph.D. Thesis, University of Texas at Austin, Graduate School of Business (1990). Also available from University Microfilms, Inc., Ann Arbor, MI.

[31] Thompson, R.G., P.S. Dharmapala and R.M. Thrall: DEA Sensitivity Analysis of Efficiency Measures with Applications to Kansas Farming and Illinois Coal Mining. In A. Charnes, W.W. Cooper, A. Lewin and L.M. Seiford, eds. *Data Envelopment Analysis: Theory, Methods and Applications*. Kluwer Academic Publisher, 1995.

[32] Thompson, R.G., R.D. Singleton, R.M. Thrall and B.A. Smith: Comparative Site Evaluations for Locating a High-Energy Physics Laboratory in Texas. *Interfaces* 16 (1986) 16-26.