Invited Review

Data envelopment analysis (DEA) – Thirty years on

Wade D. Cook a,*, Larry M. Seiford b

a Department of Operations Management and Information Systems, Schulich School of Business, York University, Toronto, Ontario, Canada M3J 1P3
b Industrial and Operations Engineering, College of Engineering, University of Michigan, Ann Arbor, MI, United States

Received 19 January 2008; accepted 22 January 2008
Available online 3 February 2008

Abstract

This paper provides a sketch of some of the major research thrusts in data envelopment analysis (DEA) over the three decades since the appearance of the seminal work of Charnes et al. (1978) [Charnes, A., Cooper, W.W., Rhodes, E.L., 1978. Measuring the efficiency of decision making units. European Journal of Operational Research 2, 429–444]. The focus herein is primarily on methodological developments, and in no manner does the paper address the many excellent applications that have appeared during that period. Specifically, attention is primarily paid to (1) the various models for measuring efficiency, (2) approaches to incorporating restrictions on multipliers, (3) considerations regarding the status of variables, and (4) modeling of data variation.

© 2008 Elsevier B.V. All rights reserved.

Keywords: DEA; Models; Multiplier restrictions; Data variation

1. Introduction

Efficiency measurement has been a subject of tremendous interest as organizations have struggled to improve productivity. Reasons for this focus were best stated fifty years ago by Farrell (1957) in his classic paper on the measurement of productive efficiency.

“The problem of measuring the productive efficiency of an industry is important to both the economic theorist and the economic policy maker. If the theoretical arguments as to the relative efficiency of different economic systems are to be subjected to empirical testing, it is essential to be able to make some actual measurements of efficiency. Equally, if economic planning is to concern itself with particular industries, it is important to know how far a given industry can be expected to increase its output by simply increasing its efficiency, without absorbing further resources.”

Farrell further stated that the primary reason that all attempts to solve the problem had failed, was due to a failure to combine the measurements of the multiple inputs into any satisfactory measure of efficiency. These inadequate approaches included forming an average productivity for a single input (ignoring all other inputs), and constructing an index of efficiency in which a weighted average of inputs is compared with output. Responding to these inadequacies of separate indices of labor productivity, capital productivity, etc., Farrell proposed an activity analysis approach that could more adequately deal with the problem. His measures were intended to be applicable to any productive organization; in other words, “from a workshop to a whole economy.” Unfortunately, he confined his numerical examples and discussion to single output situations, although he was able to formulate a multiple output case.

Twenty years after Farrell’s seminal work, and building on those ideas, Charnes et al. (1978), responding to the need for satisfactory procedures to assess the relative efficiencies of multi-input multi-output production units, introduced a powerful methodology which has subsequently been titled data envelopment analysis (DEA). The original idea behind DEA was to provide a methodology whereby, within a set of comparable decision making units (DMUs), those exhibiting best practice could be identified, and would form an efficient frontier.
Furthermore, the methodology enables one to measure the level of efficiency of non-frontier units, and to identify benchmarks against which such inefficient units can be compared.

Since the advent of DEA in 1978, there has been an impressive growth both in theoretical developments and applications of the ideas to practical situations. The purpose of the current paper is to provide a sketch of the major directions in methodological developments (as opposed to a discussion of applications), in this important field during the past three decades. The coverage is by no means complete as the volume of literature is enormous, and beyond the current scope.

Section 2 reviews the various DEA models, including those that go beyond the usual definition of DEA, specifically the free disposal hull (FDH) model, cross evaluation, and minimum distance models. Section 3 goes beyond the single level models of Section 2 and examines multilevel models. Section 4 discusses various forms of multiplier restrictions used to constrain the frontier. In Section 5 the status of different types of variables is reviewed. These include non-discretionary, non-controllable, categorical, and ordinal variables. As well, we consider the issue regarding uncertainty as to the input versus output status of variables. Data variation is explored in Section 6. This includes sensitivity analysis, probability-based models, window analysis, Malmquist models for capturing times series impacts on efficiency, and statistical inference issues surrounding the efficient frontier. Concluding remarks follow in Section 7.

2. The models

2.1. The constant returns to scale (CRS) model

Consider a set of $n$ DMUs, with each DMU $j$, ($j=1,\ldots,n$) using $m$ inputs $x_{ij}$ ($i=1,\ldots,m$) and generating $s$ outputs $y_{rj}$ ($r=1,\ldots,s$). If the prices or multipliers $u_i, v_r$ associated with outputs $r$ and inputs $i$, respectively, are known, then borrowing from conventional benefit/cost theory, one could express the efficiency $e_{ij}$ of DMU$_j$ as the ratio of weighted outputs to weighted inputs, i.e.

$$
e_{ij} = \frac{\sum_r u_r y_{rj}}{\sum_i v_i x_{ij}}.$$

This benefit/cost ratio is, of course, the basis for the standard engineering ratio of productivity.

In the absence of known multipliers, Charnes et al. (1978) proposed deriving appropriate multipliers for a given DMU by solving a particular non-linear programming problem. Specifically, if DMU$_0$ is under consideration, the Charnes et al model for measuring the technical efficiency of that DMU is given by the solution to the fractional programming problem:

$$
e_0 = \max \sum_r u_r y_{ro}/\sum_i v_i x_{io}$$

s.t. \[ \sum_r u_r y_{rj} - \sum_i v_i x_{ij} \leq 0, \quad \forall j \]

\[ u_r, v_i \geq 0, \quad \forall r, i. \]

where \( \epsilon \) is a non-archimedean value designed to enforce strict positivity on the variables. We point out that this model involving the ratio of outputs to inputs is referred to as the input-oriented model. One could, as well, invert this ratio and solve the corresponding output-oriented minimization problem. We will generally deal with the input-oriented model herein.

Problem (2.1) is referred to as the CCR (Charnes, Cooper and Rhodes) model, and provides for constant returns to outputs. Data variation is explored in Section 6. This analysis, the Charnes et al model for measuring the CCR model.

Applying the Charnes and Cooper (1962) theory of fractional programming, making the change of variables $u_r = tu_r$ and $v_i = tv_i$, where $t = (\sum y_{ro})^{-1}$, problem (2.1) can be converted to the linear programming (LP) model:

$$e_0 = \max \sum_r \mu_r y_{ro}$$

s.t. \[ \sum_i v_i x_{io} = 1 \]

\[ \sum_r \mu_r y_{rj} - \sum_i v_i x_{ij} \leq 0, \quad \forall j \]

\[ \mu_r, v_i \geq \epsilon, \quad \forall r, i. \]

By duality, this problem is equivalent to the linear programming problem:

$$\min \theta_0 - \epsilon \left( \sum_r s^r + \sum_i s^i \right)$$

s.t. \[ \sum_j \lambda_j x_{ij} + s^i = \theta_0 x_{io}, \quad i = 1, \ldots, m \]

\[ \sum_i \lambda_i y_{rj} - s^r = y_{ro}, \quad r = 1, \ldots, s \]

\[ \lambda_j, s^r, s^i \geq 0, \quad \forall i, j, r \]

\[ \theta_0 \text{ unconstrained}. \]

Problem (2.3) is referred to as the envelopment or primal problem, and (2.2) the multiplier or dual problem.

The constraint space of (2.3) defines the production possibility set $T$. That is,

$$T = \left\{ (X, Y) \mid X \geq \sum_j \lambda_j X_j, Y \leq \sum_j \lambda_j Y_j, \lambda_j \geq 0 \right\}.$$

To get a geometric appreciation for the CRS model, one can represent problem (2.3) in a form such as pictured in Fig. 1.

This figure provides an illustration of a single output single input case. If we solve (2.3) for each of the DMUs, this amounts to projecting that DMU to the left, to a point on the frontier. In the case of DMU #3, for example, its
the resulting projected value $\theta'_{EX}$ is simply the frontier DMU B. We may refer to B as a “benchmark” for DMU E. In the case of DMU G, its projected (frontier) value is represented by the point K, hence B and C are appropriate benchmarks for DMU G. As illustrated by this example, the CCR model is appropriately referred to as providing a radial projection. Specifically, each input is reduced by the same proportionality factor $\theta$.

In the example of Fig. 2, in solving (2.3) for DMUs A, B, C, D, E, F, G, all slack variables (in this case $s_1^i, s_2^i$) will be zero. Fig. 3 is a redrawn version of Fig. 2 with an additional DMU H added. For the DMU H, however, the projected point lies on an extension of the frontier at H*, and not on the frontier proper. In this case, the slack corresponding to input #1 (s1) will be positive, and equal to the distance represented by the line segment from H* to D. We say that DMU H is improperly enveloped. A DMU located at point H* would be deemed weakly efficient, as opposed to points such as A, B, C, D which are strongly efficient or simply efficient (DEA-efficient). For a complete discussion of efficiency classes, the reader is referred to Charnes et al. (1986, 1991).

2.2. The variable returns to scale (VRS) model

Banker et al. (1984) (BCC), extended the earlier work of Charnes et al. (1978) by providing for variable returns to scale (VRS). This is pictured in the redrawn version of Fig. 4 in the form of Fig. 4.

Shown are the original CRS frontier, and the VRS frontier, here represented by the line segments 1–2, 2–3 and 3–4. The BCC ratio model differs from (2.1), by way of an additional variable, i.e.

$$e^*_o = \max \left\{ \sum_{i} u_i x_{ro} - u_o \right\} / \sum_{i} v_i x_{io}$$

s.t. $\sum_{i} u_i x_{io} - u_o - \sum_{i} v_i x_{ij} \leq 0, \quad j = 1, \ldots, n$$

$u_i \geq e, \quad v_i \geq e, \quad \forall i, r$

$u_o$ unrestricted in sign.

$$\sum_{i} v_i x_{ij} \leq 0, \quad j = 1, \ldots, n$$

$$u_i \geq e, \quad v_i \geq e, \quad \forall i, r$$

$u_o$ unrestricted in sign.

$$\sum_{i} v_i x_{ij} \leq 0, \quad j = 1, \ldots, n$$

$$u_i \geq e, \quad v_i \geq e, \quad \forall i, r$$

$u_o$ unrestricted in sign.

$$\sum_{i} v_i x_{ij} \leq 0, \quad j = 1, \ldots, n$$

$$u_i \geq e, \quad v_i \geq e, \quad \forall i, r$$

$u_o$ unrestricted in sign.

$$\sum_{i} v_i x_{ij} \leq 0, \quad j = 1, \ldots, n$$

$u_i \geq e, \quad v_i \geq e, \quad \forall i, r$
The linear programming equivalent of (2.4) is
\[
e^o_o = \max \sum \mu_i y_{ro} - \mu_o
\]
s.t. \( \sum v_i x_{io} = 1 \)
\[
\sum \mu_i y_{rj} - \mu_o - \sum v_j x_{io} \leq 0, \quad j = 1, \ldots, n
\]
\[
\mu_r \geq e, \quad v_j \geq e, \quad \forall i, r, \quad \mu_o \text{ unrestricted.}
\]
(2.5)

It is noted that (2.6) differs from (2.3) in that it has the additional convexity constraint on the \( \lambda_j \), namely \( \sum \lambda_j = 1 \).

In reference to Fig. 4, that portion of the frontier from point 1 up to (but not including) point 2, constitutes the increasing returns to scale portion of the frontier; point 2 is experiencing constant returns to scale; all points on the frontier to the right of 2 (i.e., the segments from 2 to 3 and from 3 to 4) make up the decreasing returns to scale portion of the frontier. As with the CRS model, a DMU is BCC-efficient in the VRS sense if there exists a solution to (2.6) such that \( \theta^*_o = 1 \) and all slacks \( s^*_r, s^*_j \) are zero in value. Clearly, any CCR-efficient DMU is also BCC-efficient.

The returns to scale (RTS) classification of DMUs has been the subject of study by numerous authors, including Banker (1984) (using the most productive scale size concept and letting the sum of lambda values dictate the RTS), Banker et al. (1984) (using the free variable in (2.5)), and Färe et al. (1994), (applying their scale efficiency index method). A problem in classifying RTS is the existence of multiple optima, meaning that the classification may be a function of the particular solution selected by the optimization software. Various attempts have been made to provide a more definitive RTS classification assignment for a given DMU, including developing intervals for the various free variables arising from the multiple optima. Zhu and Shen (1995) suggest a remedy for the CCR RTS method under multiple optima. Seiford and Zhu (1997, 1999b) review the various methods and suggest computationally simple methods to characterize RTS, and to circumvent the need for exploring all alternate optimal solutions.

2.3. The additive model

The previous two efficiency models are radial projection constructs. Specifically, in the input-oriented case, inputs are proportionally reduced while outputs remain fixed. (For the output-oriented case, outputs are proportionally increased while inputs are held constant.) Charnes et al. (1985b) introduced the additive or Pareto–Koopmans (PK) model which, to an extent, combines both orientations. Fig. 5 illustrates this idea wherein any direction in the quadrant formed by B–A–C is permitted.
There are several versions of the additive model, the most basic being given by the linear optimization problem shown as (2.7). The convexity condition on the $\lambda_j$ variables implies that we are using the VRS technology. The frontier generated by model (2.7) is identical to that arising from the corresponding VRS structure (2.6), hence a DMU is additive-efficient or PK efficient (all slacks equal to zero at the optimum in (2.7)) if and only if it is VRS-efficient. Clearly, the CRS production possibility set can be used as well (and is the one illustrated in Fig. 5).

$$P_o = \max \sum_i s_i^- + \sum_r s_r^+$$

s.t. $\sum_j \lambda_j x_{ij} + s_i^- = x_{io}, \quad i = 1, \ldots, m$

$$\sum_j \lambda_j y_{ij} - s_i^+ = y_{io}, \quad r = 1, \ldots, s$$

(2.7)

$$\sum_j \lambda_j = 1$$

$\lambda_j, s_i^-, s_i^+ \geq 0, \quad \forall j, i, r.$

Since the various inputs and outputs may be measured in non-commensurate units, (Russell, 1988), it may not be practical in certain contexts to use the simple sum of slacks as the objective in (2.7). Moreover, model (2.7) does not provide for an actual measure of inefficiency as in the case for the BCC and CCR models. To overcome this latter problem, Charnes et al. (1985b) proposed the use of $Q_o$, where

$$Q_o = \delta \left( \sum_i s_i^- / x_{io} + \sum_r s_r^+ / y_{ro} \right)$$

subject to the constraints as in (2.7). A suggested value for $\delta$ was 1/(m+s). The division of the $s_i^-$ and $s_r^+$ by $x_{io}$ and $y_{ro}$, respectively, is intended to render these slacks units invariant (i.e., commensurate), while multiplying by $\delta$ controls the overall scale. In order to maintain consistency with the sense of efficiency in the CCR and BCC models, Sueyoshi (1990) offers $1 - Q_o$ as such a measure. The problem, as acknowledged in a later paper by Chang and Sueyoshi (1991), is that $0 \leq 1 - Q_o \leq 1$ may not necessarily hold, and may in fact be negative.

2.4. Slacks-based measures

To address the above shortcomings in the additive model, Green et al. (1997) propose as a measure of efficiency

$$R_o = \frac{1}{s+r} \left[ \sum_i s_i^- / x_{io} + \sum_r s_r^+ / (y_{ro} + s_r^+) \right],$$

and recommend solving the problem

$$\max \ R_o$$

Subject to the constraints of (2.7).

While the non-linearity of $R_o$ poses a computational inconvenience, the resulting measure $1 - R_o$ does possess the property of being on the unit scale [0,1], hence serving as a legitimate efficiency score.

Tone (2001) introduced the so-called slacks-based measure (SBM) which is invariant to the units of measurement and is monotone increasing in each input and output slack. The SBM is derived from the solution of the fractional programming problem

$$\min p = \frac{1 + \sum s_i^- / x_{io}}{1 + \sum s_r^+ / y_{ro}}$$

Subject to the constraints of (2.7).

Clearly, $0 \leq p \leq 1$, and is therefore a legitimate PK efficiency score in the spirit of the CCR and BCC models. It is shown in Tone (2001) that (2.9) can be transformed into a linear programming problem.

2.5. The Russell measure

The Russell measure model, named by Färe and Lovell (1978), and later revisited by Pastor et al. (1999) (referring to it as the enhanced Russell measure), is equivalent to Tone’s SBM, as discussed in Cooper et al. (2006). Specifically, the model is

$$R_o = \min \left[ \sum_i (0. / m) / \sum_r (q_r / s) \right]$$

s.t. $\sum_j \lambda_j x_{ij} \leq \theta x_{io}, \quad i = 1, \ldots, m$

$$\sum_j \lambda_j y_{ij} \geq \varphi y_{ro}, \quad r = 1, \ldots, s$$

$$\sum_j \lambda_j = 1$$

$\lambda_j \geq 0, \quad 0 \leq \theta \leq 1, \quad \varphi \geq 1, \quad \forall i, j, r.$

(2.10)

2.6. Other non-radial models

There are several other non-radial models. One of these is the RAM (range adjusted measure) model of Cooper et al. (1999a), and is similar to the additive model with the additional feature that the score lies on the [0,1] scale. There are also non-radial models employed as a second stage in a two-stage efficiency analysis, after a projection point has been identified for a given DMU. See Tone (2001), Cooper et al. (2001), Portela and Thanassoulis (2007), Portela et al. (2003), and others.

2.7. Alternative views

2.7.1. FDH – The free disposal Hull model

A production possibility set (PPS) or reference technology can be thought of as a declaration of the totality of production activities that might plausibly have been observed on the evidence of the activities actually observed.
DEA uses the frontier of the PPS, defined in terms of observed activities deemed efficient, and specific linear convex combinations thereof, to evaluate the observed activities. Deprins et al. (1984) and Tulkens (1993) take an alternate view, whereby the assumption is that only the observed DMUs make up the frontier, not linear or convex combinations of those observed units.

The model that will generate this frontier is simply the VRS model of Banker et al. (1984), but with the additional restriction that the \( \lambda_j \in [0,1] \). While Thrall (1999) challenged this concept, from an economic theory perspective, this was strongly rebutted by Cherchye et al. (2000), and FDH remains an attractive, but potentially underutilized approach to efficiency measurement. See also Green and Cook (2004).

### 2.7.2. Cross efficiency

The cross efficiency score of a given DMU is obtained by computing for that DMU the set of \( n \) technical efficiency scores (using the \( n \) sets of optimal weights corresponding to the \( n \) DMUs), and then averaging those scores. Thus, cross efficiency goes beyond pure self-evaluation inherent in conventional DEA analysis, and combines this with the other \( (n-1) \) scores arising from the optimal peer multipliers. This approach was originated by Sexton et al. (1986), and was further investigated by Doyle and Green (1994), and others. Cross efficiency provides an efficiency ordering among all the DMUs to differentiate between good and poor performers. It eliminates the need for incorporating additional weight restrictions on multipliers, thereby avoiding potentially unrealistic weighting schemes (Anderson et al., 2002). One can find many uses of cross efficiency for example, R&D project selection (Oral et al., 1991), preference voting (Doyle et al., 1996) and others. As pointed out by Doyle and Green (1994), the non-uniqueness of the DEA optimal weights possibly reduces the usefulness of cross efficiency. To combat this problem, those authors have proposed various secondary goals such as given by the aggressive and benevolent models. Liang et al. (2008) further improve on the idea of the cross efficiency score, using game theoretic constructs. To implement their idea, one views DMUs as players in a game, and defines the efficiency score as that arising from (2.1).

In a sense, suppose one player DMU \( j \) is given an efficiency score \( x_{ij} \), and that another player DMU, then tries to maximize its own efficiency, given that \( x_{ij} \) cannot be decreased. The authors present an algorithm that continually updates the \( x_{ij} \) arriving at a final set of scores that are, in a competitive sense, best for the set of DMUs.

### 2.8. Least distance projections

A number of authors have examined the problem of deriving the least distance projection to the efficient frontier; note that this is the opposite criterion to that of the additive model that searches for the greatest distance. Frei and Harker (1999) proposed using the Euclidean norm to define the closest point. Charnes et al. (1992, 1996) and Bries (1999) obtain the minimum city block distance to the weak efficient frontier. Gonzalez and Alvarez (2001) minimize input contractions, while Portela et al. (2003), Cherchye and Van Puyenbroeck (2001), approach this by identifying all the efficient facets. In a recent paper Aparicio et al. (2007) present a set of models for obtaining least distance projections.

### 2.9. Invariance to data alterations

Either out of necessity or convenience, the modeler is sometimes called upon to alter or transform the data to be used in a DEA analysis. For example, it may be more convenient for scale purposes to represent a resource in thousands of dollars rather than in dollars. If certain profit figures can take negative values it may be desirable to translate the data by adding a fixed number to the value of that variable for each DMU (thereby rendering all values positive). An important consideration is whether or not such alterations made to original data, influence the outcomes arising from the application of the various efficiency measurement models discussed above. This question of invariance has been a subject of importance in the DEA literature, and is discussed, for example, in Ali and Seiford (1990), Thrall (1996), Pastor (1996) and Cooper et al. (2006).

In the case of a given factor (e.g., \( x_{ij} \)), two specific forms of data alteration are of particular significance:

<table>
<thead>
<tr>
<th>Model</th>
<th>CCR-I</th>
<th>CCR-O</th>
<th>BCC-I</th>
<th>BCC-O</th>
<th>ADD</th>
<th>SBM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Trans.</td>
<td>Y</td>
<td>Free</td>
<td>Free</td>
<td>Free</td>
<td>Semi-p</td>
<td>Free</td>
</tr>
<tr>
<td>Invariance</td>
<td>-Y</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>Units invariance</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>( \rho )</td>
<td>([0,1])</td>
<td>([0,1])</td>
<td>([0,1])</td>
<td>No</td>
<td>([0,1])</td>
<td></td>
</tr>
<tr>
<td>Tech. or mix</td>
<td>Tech.</td>
<td>Tech.</td>
<td>Tech.</td>
<td>Tech.</td>
<td>Mix</td>
<td>Mix</td>
</tr>
<tr>
<td>Returns to scale</td>
<td>CRS</td>
<td>CRS</td>
<td>VRS</td>
<td>VRS</td>
<td>C(V)RS(^a)</td>
<td>C(V)RS</td>
</tr>
</tbody>
</table>

\(^a\) The Additive model is translation invariant only when the convexity constraint is added.
\(^b\) C(V)RS means Constant or Variable returns to scale according to whether or not the convexity constraint is included.
3. Multistage/serial models

3.1. Network DEA

This body of work originated by Fare and Grosskopf (1996) is built around the concept of sub-technologies within the “black box” of DEA. This approach allows one to examine in more detail the inner workings of the production process, potentially leading to a better understanding of that process. Färe et al present three general network models:

1. A static network model in which a finite set of sub-technologies or activities are connected to form a network. Its main feature is that it allows one to analyze the allocation of intermediate products. Fig. 6 illustrates the concept.
2. Dynamic network model: This structure permits one to examine a sequence of production technologies where a decision at one stage (e.g. a time period) impacts later stages. Here, intermediate products are accounted for, meaning that the outputs of one stage become inputs to a later stage.
3. Technology adoption: This model allows one, for example, to examine production on different processors (e.g. machines). Inputs are allocated among the processors to allow one to determine which technology to adopt.

3.1.2. Supply chains

Several DEA-based approaches have been used to examine buyer–supplier supply chain settings, leading to efficiency evaluation. The important issue is that of deriving a measure of overall efficiency as opposed to looking only at the efficiencies of individual members of the chain. Therein, of course, lies the difficulty in that a recommended efficiency improvement for one member of the chain can lead to a decrease in the efficiency of the other members. Seiford and Zhu (1999c) and Chen and Zhu (2004) provide two approaches in modeling two-stage processes. Zhu (2003b) presents a DEA-based supply chain model to both measure the overall efficiency, and that of its members. Fig. 7 captures the type of structure examined by Zhu (2003b).

A number of supply chain approaches due to Liang et al. (2006) are built on game theoretic constructs. Liang’s two principal models are (1) a non-cooperative model and (2) a cooperative model. In the non-cooperative model, he views the seller as the leader and buyer as the follower. In the first stage, one optimizes the leader’s efficiency score and then maximizes (in the second stage) that of the follower, with the constraint that the multipliers used must be such that the first stage (leader) score remains unchanged. The resulting model is a non-linear parametric programming problem. In the cooperative game model, no leader – follower assumption is made.

3.2. Multicomponent/parallel models

The multilevel settings of the previous subsection are generally directed toward what can be termed serial processes. Some situations can involve multiple components that can be regarded as operating in parallel. In Cook et al. (2000) a study of bank branch performance is discussed but where in each branch activities can be grouped under two headings – sales activities and service activities. While one could conceivably consider looking at the two sets of activities as involving separate analyses, the complication that arises is that of dividing up the shared inputs such as support staff. Cook et al. (2000) instead propose an aggregate efficiency measurement.
model that allows one to evaluate both the component level efficiencies as well as the overall or aggregate efficiency. Another approach to multicomponent situations in organizations is given by Portela et al. (2007).

3.3. Hierarchical/nested models

Some multilevel efficiency measurement situations can involve hierarchical or nested structures. Cook et al. (1998), Cook and Green (2005) examine a set of power plants, wherein each plant is made up of a set of individually operating power units. At one level, it is necessary to consider measuring the efficiency of each power unit relative to the full set of power units across all plants; here, the power units are the DMUs. At the same time, at the next level up in the hierarchy, one can evaluate the efficiency of each plant or group of units, against all other plants. There can be inputs and outputs at one level that are not part of the analysis at another level. Cook et al. (1990) give another example of such a structure in considering the efficiency of highway maintenance crews or patrols. Here, patrols (level 1 DMUs) are grouped under districts (level 2 DMUs), which are further grouped under the different regions (level 3 DMUs) within the province or state.

4. Multiplier restrictions

Within the DEA literature, there is a body of research aimed at addressing the problem of unacceptable weighting schemes. We refer to the collection of methodologies here as involving multiplier restrictions, although as discussed below, some of the tools do this in an indirect rather than direct way (e.g., via constrained facet analysis).

4.1. Absolute multiplier restrictions

Some of the earliest work here involved imposing absolute lower and upper bounds on input and output multipliers, that is

\[ P_{1r} \leq \mu_r \leq P_{2r}, \quad Q_{1i} \leq \nu_i \leq Q_{2i}. \]

Roll et al. (1991) examined the use of such absolute limits in the context of evaluating highway maintenance units (see also Cook et al. (1990)). In an earlier paper by Dyson and Thanassoulis (1988), a similar approach was proposed. Implementing absolute bounds can prove difficult, in that appropriate levels for the \( P_{kr}, Q_{ki} \) are very much a function of the scales used for the variables. Only after running an “unbound” model will the range of possible multiplier values be known in relation to the scales adopted.

4.2. Cone ratio restrictions

Charnes et al. (1990), in their study of large industrial banks, recognized that undesirable weighting schemes are a natural outcome in many DEA applications. To provide for more realistic multipliers, they proposed imposing a set of linear restrictions that define a convex cone. Specifically, the feasible region for say the input multiplier vector \( v = (v_1, \ldots, v_m) \) is defined to be in the polyhedral convex cone spanned by a set of \( k \) admissible non-negative direction vectors \( (a_i), \ell = 1, \ldots, k \). Thus, a feasible \( v \) can be expressed as

\[ v = \sum_{\ell} x_{\ell}a_{\ell}, \quad x_{\ell} \geq 0, \quad \forall \ell. \]

Let the resulting polyhedral cone be denoted by \( V \), the set of all \( v \) satisfying (3.1). Letting \( U \) be a similar cone defining the set of feasible output multiplier vectors \( \mu = (\mu_1, \ldots, \mu_s) \), the CCR cone ratio model is then given by

\[
\begin{align*}
\max & \quad \sum_{r} \mu_r y_{ro} \\
\text{s.t.} & \quad \sum_{i} v_i x_{i0} = 1 \\
& \quad \sum_{r} \mu_r y_{rij} - \sum_{i} v_i x_{ij} \leq 0, \quad j = 1, \ldots, n \\
& \quad \mu \in U, \quad \mu \in U.
\end{align*}
\]

There is, of course, the corresponding dual of this model which can be found in Charnes et al. (1990).

An important generalization in (3.2) is that given by Thompson et al. (1995), wherein the spaces for \( \mu \) and \( v \) are connected by way of “linked” cones. Those authors apply their general cone ratio model to a problem involving Illinois coal mining.

4.3. Assurance regions

A special case of the cone ratio idea is what Thompson et al. (1986, 1990) termed an assurance region (AR). The AR concept was developed to prohibit large differences in the values of multipliers, and imposes constraints on the relative magnitudes of those multipliers. For example, one might add a constraint on the ratio of multipliers for a pair of inputs 1 and 2, in the form:

\[ L_{12} \leq v_2/v_1 \leq U_{12}, \]

where \( L_{12}, U_{12} \) are lower and upper bounds, respectively, on the ratio \( v_2/v_1 \).

Generally, the imposition of multiplier restrictions, whether it is through absolute bounds, cone ratio constraints, or AR constraints, leads to a worsening of efficiency scores. Referring again to Fig. 2, a redrawn version is shown in Fig. 2’. The slopes of the facets in this figure are related to the relative values of \( v_1 \) and \( v_2 \). When restrictions say of the form (3.3) are imposed on the DEA model, certain slopes may no longer be admissible. This has the affect of “bending” the frontier out (giving it less curvature), as illustrated by the dashed line in Fig. 2’. Thus, DMUs that are efficient in an unrestricted setting, e.g. DMU D in Fig. 2, may be rendered inefficient as in Fig. 2’. In their recent book, Cooper et al. (2006), p. 172 provide a useful alternative pictorial view of this idea in the multiplier space.

Many applications of the AR form of the various DEA models can be found in the literature. These help to
enlighten the reader on the practicalities of deriving appropriate bounds. In some cases, these bounds are presented by the author as being “illustrative” only, and one is left with the question as to how the “right” values could be derived. In other circumstances, bounds may fall naturally out of the available data. In Cook et al. (2000), for example, in studying bank branch efficiency, outputs are various classes of branch transactions (deposits, withdrawals, etc.), and their multipliers are transaction processing times, in minutes or hours. While exact times are not given, since these can vary from branch to branch, from one employee to another etc., ranges with established lower and upper limits are available. It is these ranges that lead to concrete limits of the form illustrated in (3.3).

Various generalizations on the AR concept appear in the literature. Allen et al. (1997) and Thanassoulis et al. (1998) present a global type of restriction on the weighted values for each DMU. For example, if input #1 is to represent at least 10%, and at most 20% of the total weighted input for any DMU, then constraints such as $0.10 \leq v_1 x_{1i} / \left( \sum v_i x_{ij} \right) \leq 0.20$ would be appropriate. Note that separate constraints for each DMU would result from this.

Cook and Zhu (2008) present a context-dependent assurance region DEA (CAR-DEA) model which provides for restrictions of the form (3.3) that may differ from one subset of DMUs to another. Thus, if there are $K$ groups of DMUs, and if we wish to impose restrictions on say outputs of the form $c_{ik}^x \leq \mu_i / \mu_r \leq c_{ik}^x$, $k = 1, \ldots, K$,

it is necessary to account for potential inconsistency (infeasibility) if all $K$ sets are imposed simultaneously. If we assume that output #1 is used as the numeraire against which other outputs are compared, the approach taken by Cook and Zhu (2007) is to replace the set of $K$ groups of AR restrictions by a single set of AR constraints, applicable to all $K$ classes of DMUs.

### 4.4. Facet models

Significant work has been done relating to facet extension and facet identification, to address the inherent problem involving the occurrence of zero weights (or $e$-weights) in the multiplier models, as indicated above. This is equivalent to projection to weakly efficient facets or non-full-dimensional facets. Bessent et al. (1988) were the first to introduce the idea of constrained facet analysis (CFA). In the event that a given unit is projected to a weakly efficient facet, CFA involves extending a selected Pareto-efficient (full-dimensional) facet, and then projecting the given DMU on to that extended facet. Lang et al. (1995) improved on this idea by adopting a two-stage approach which ultimately amounts to finding the “closest” full-dimensional facet to which to project the DMU in question. Other similar approaches have been suggested by Green et al. (1996), and by Olesen and Petersen (1996).

As an example, the method of Green et al. (1996) is a three-stage procedure. In stage 1, the standard CCR model is solved to determine the full set of efficient units $E \cup E'$. Note that $E$ is composed of extreme efficient DMUs (corners of facets) while $E'$ are the non-extreme (interior to facets) efficient units. Stage 2 then follows Charnes et al. (1986) to partition this full set of units into the two subsets. Finally, in the third stage one solves the mixed integer programming problem,

$$\max \ e_0 = \sum_{r} \mu_r y_{re}$$
$$\text{s.t.} \quad \sum_{r} \mu_r y_{re} - \sum_{i} v_i x_{ie} \leq 0, \ ee E,$$
$$\quad \sum_{r} \mu_r y_{re} - \sum_{i} v_i x_{ie} + M z_e \geq 0, \ ee E,$$
$$\quad \sum_{i} v_i x_{ie} = 1,$$
$$\quad \sum_{i} z_e = | E | - (m + s - 1),$$
$$\quad \mu_r \geq 0, r = 1, \ldots, s, \quad v_i \geq 0, i = 1, \ldots, m,$$
$$\quad z_1 \in \{0, 1\} ee E, \quad M \gg 0.$$

This model guarantees that exactly $m + s - 1$ of the $m + s$ constraints will be satisfied as equalities. This means that at most one slack variable will be positive, hence all $\mu_r, v_i$ variables will be forced to be strictly positive, meaning that the DMU in question is projected against a full-dimensional facet. The method is illustrated by Fig. 8.

#### 4.5. Generating unobserved DMUs

It is noted that in extending facets to eliminate weakly efficient projections, new “unobserved” DMUs will be generated. In Fig. 8, points where the rays from the origin to improperly enveloped DMUs ($P_d$ and $P_r$) intersect the extended facet, define such DMUs. Thanassoulis and Allen (1998) present a formal procedure for producing new unobserved DMUs, thus creating the means for extending observed facets. Their approach amounts to obtaining
information from the decision maker as to his/her estimates of the tradeoff between pairs of factors.

Another line of research in a similar direction is due to Golany and Roll (1994) who introduced the idea of incorporating standard DMUs into the DEA structure. Cook and Zhu (2005) extended this work by way of incorporating production standards (as opposed to standard DMUs).

5. Special considerations regarding the status of variables

As originally conceived, the DEA model involves the generation of \( s \) outputs \( \{y_{ij}\}_{j=1}^{s} \) using \( m \) inputs \( \{x_{ij}\}_{j=1}^{m} \). In a structure such as that in (2.3) all inputs are projected radially to the efficient frontier, all variables are assumed to be quantitative, and the collection of DMUs under evaluation is assumed to form a relatively homogeneous group (all are comparable to one another). As new applications of DEA have arisen, it has become necessary to expand the original model structures to accommodate new situations, hence relaxing a number of those original assumptions. In this section we discuss some of the more important developments in this regard.

5.1. Non-discretionary variables

In many applications of DEA, certain of the input variables may not be under the direct control of management. In a DEA analysis of bank branch efficiency, for example, an input variable such as fixed expenditures (rent, utilities, etc.) could not be proportionally reduced as would be the case for variable expenditures such as staff. Thus, it is important to identify those variables that are discretionary (staff) versus non-discretionary (fixed costs).

Banker and Morey (1986a) introduced the first DEA model that allowed for non-discretionary inputs by modifying the input constraints to disallow input reduction on the fixed factor. Letting \( D \) denote the subset of inputs \( i \in \{1, 2, \ldots, m\} \) that are discretionary, and ND, the non-discretionary inputs, the Banker and Morey model becomes

\[
\min \theta_0 - \varepsilon \left( \sum_{i \in D} s_i^+ + \sum_{r} s_r^k \right) \\
\text{s.t.} \ \sum_{j} \lambda_j x_{ij} + s_i^+ = \theta_0 x_{io}, \ i \in D \\
\sum_{j} \lambda_j y_{ij} - s_i^- = y_{io}, \ i \in ND \\
\lambda_j, s_i^+, s_i^- \geq 0, \ \forall i, j, \ \theta_0 \text{ unrestricted.}
\]

The corresponding dual problem is

\[
\max \sum_r \mu_r y_{ro} - \sum_{i \in ND} v_i x_{io} \\
\text{s.t.} \ \sum_r \mu_r y_{rj} - \sum_{i \in D} v_i x_{ij} - \sum_{i \in ND} v_i x_{io} \leq 0, \ j = 1, \ldots, n \\
\mu_r \geq \varepsilon \forall r, \ v_i \geq \varepsilon \forall i \in D, \ v_i \geq 0, i \in ND.
\]

It is noted that for non-discretionary \( i \in \text{ND}, \ v_i \geq 0 \) rather than \( v_i \geq \varepsilon \). Correspondingly, in (4.1) it is only those input slacks related to discretionary factors that appear in the objective function. See Cooper et al. (2006), pp. 210–211 for a full discussion of this.

More recently Ruggiero (1996) has pointed out that in certain cases, the Banker and Morey model can over estimate technical efficiency by allowing production impossibilities into the referent set. Ruggiero’s approach restricts weights to zero for production possibilities with higher levels of the non-discretionary inputs, and as a result, production impossibilities are appropriately excluded from the referent set. Recent simulation analyses in Syrjanen (2004) and Muniz et al. (2006) demonstrate that Ruggiero’s method performs relatively well in evaluating efficiency. See also Ruggiero (1998, 2007).

5.2. Non-controllable variables

In the non-discretionary variable model (4.1) we note that the slacks \( s_i^-, i \in \text{ND} \) are permitted to be strictly positive. This means that at the optimum, a DMU \( u \) may end up being compared to a linear combination of peers wherein the value of a non-discretionary variable for that combination is less than its value in DMU \( u \). In certain situations there can be inputs (and outputs) whose values must remain fixed, and can only be compared to DMUs whose linear combinations are at the same levels as those fixed factors. Such variables have been labeled as non-controllable. Let \( N_1, N_2 \) denote non-controllable inputs and outputs respectively, and \( N_1, N_2 \) denote regular (controllable) inputs and outputs, respectively. The non-controllable variable model is then:

\[
\min \theta_0 - \varepsilon \left( \sum_{i \in N_1} s_i^+ + \sum_{r \in N_2} s_r^k \right) \\
\text{s.t.} \ \sum_{j} \lambda_j x_{ij} + s_i^+ = \theta_0 x_{io}, \ i \in N_1 \\
\sum_{j} \lambda_j y_{ij} - s_i^- = y_{io}, \ i \in N_1 \\
\sum_{j} \lambda_j y_{rj} - s_r^+ = y_{ro}, \ r \in N_2 \\
\lambda_j \geq 0, \ \forall j; \ s_i^- \geq 0, \ i \in N_1, s_i^+ \geq 0, \ r \in N_2.
\]

5.3. Categorical variables (categorical DMUs)

There are situations in which DMUs fall into natural categories. An example would be when we are evaluating a set of retail establishments wherein different levels of service exist from one establishment to another. To provide a fair evaluation of each DMU, it can be argued that a DMU in any given category should be compared only to those other units in the same or less-advantaged
categories. A DMU under heavy competition would be unfairly penalized if compared to units in significantly more favorable competitive environments.

Banker and Morey (1986b) presented the first model to deal with such situations. Their model saw the introduction of categorical inputs \( x_{ij}, i = m + 1, \ldots, m \) in addition to regular controllable inputs \( x_{ij}, i = 1, \ldots, m' \). In the case where a categorical variable \( x_{ij} \) is not controllable by management, their treatment involves replacing that variable by a set of binary variables \( d_{ij}^{(k)}, k = 1, \ldots, K_i \), with \( K_i \) being the number of categories. Arranging the categories in decreasing order of favorability one sets \( d_{ij}^{(k)} = 1, k \leq k_0 \), and \( d_{ij}^{(k)} = 0, k > k_0 \), if DMU \( i \) is in category \( k_0 \). The usual non-discretionary input constraint \( \sum_{j} d_{ij}^{(k)} x_{ij1} + s_{j}^{+} = x_{io} \) is then replaced by a set of \( K_i \) constraints \( \sum_{j} d_{ij}^{(k)} x_{ij1} \leq d_{io}^{(k)} \).

In this way, a DMU is compared only to other DMUs in the same or less favorable categories.

In the case of a controllable categorical variable, Banker and Morey (1986b) presented a mixed integer LP formulation. However, as pointed out by Kamakura (1988), the Banker and Morey model was flawed due to a mis-specified constraint, and a revised model was given. The Kamakura model presented its own difficulties, however, and these were addressed in a later paper by Rousseau and Semple (1993). The latter authors dealt with both input and output categorical variables, and were able to reduce the integer problem to a more conventional LP approach.

5.4. Ordinal variables/data

DEA analyses are generally based on a set of quantitative input and output factors. In certain settings, however, qualitative variables may be present. For a factor such as management competence, for example, one may be able to provide only a ranking of the DMUs from best to worst. The capability of providing a more precise, quantitative measure reflecting such a factor is often not feasible. In some situations such factors can be “quantified”, but often such quantification is superficially forced as a modeling convenience.

The original DEA models incorporating rank order variables are due to Cook et al. (1993, 1996). To capture such rank order variables within the DEA structure, the authors proceed as follows. For each output \( r \), for example, assume a DMU \( k \) can be assigned to one of \( L \) rank positions \( (L \leq n) \). One can view the assignment of DMU \( k \) to position \( \delta \) on ordinal output \( r \), as having assigned that \( k \) an output value or worth \( y_{r(\delta)} \).

More recently, Cooper et al. (1999b) examined the DEA structure in the presence of what they termed imprecise data (IDEA). Zhu (2003a) and others have extended the Cooper et al. (1999a) model. While various forms of imprecision are looked at under the umbrella of IDEA, the principal focus is on rank order data. In a recent paper by Cook and Zhu (2006), rank order variables and IDEA are revisited, and both discrete and continuous projection models are discussed. It is shown that the IDEA approach for rank data is equivalent to the Cook et al. (1993, 1996) methodology.

5.5. Modelling undesirable factors

The usual variables in DEA are such that more is better for outputs, and less is better for inputs. In some situations, however, a factor can behave opposite to this; consider, for example, air pollution as one of the outputs from power plants. A number of authors have addressed this issue, in particular, Scheel (2001), Seiford and Zhu (2002), Fare and Grosskopf (2004) and Hua and Bin (2007). Approaches range from linear transformations of the original data to the use of directional distance functions.

5.6. Flexible measures – Classifying inputs and outputs

In the standard applications of DEA it is assumed that the input versus output status of each performance measure related to the DMUs is known. In some situations, however, the role of a variable may be flexible. Consider the example of measuring power plant efficiency as discussed in Cook et al. (1998) and Cook and Green (2005), where one of the outputs is a function of what is termed ‘outages.’ This measure is designed to represent the percentage of time that a plant is available to be in operation, and can, therefore, be viewed as a type of accomplishment (output) on the part of management. At the same time, it is reasonable to view this variable as an environmental input that has a direct influence on plant performance.

The incorporation of such flexible variables into the DEA structure present a problem in that there is a need to make allowance for them on both the input and output sides of the model. Beasley (1995) dealt with a similar problem, and presented a formulation for a situation where the variable ‘research funding’ was counted as \( \text{both} \) an input and an output in evaluating universities in the UK. Cook et al. (2006) later showed that Beasley’s model was flawed and gave an alternative, corrected version. The flexible variable problem is not one of counting the influence in both places, but rather counting it in the most appropriate place.

Suppose there exist \( L \) flexible measures, whose input/output status we wish to determine. Denote the values assumed by these measures as \( w_{ij}(\ell) \) for DMU \( j (\ell = 1, \ldots, L) \). For each measure \( \ell \), introduce the binary variables \( d_{ij} \in \{0,1\} \), where \( d_{ij} = 1 \) designates that factor \( \ell \) is an output, and \( d_{ij} = 0 \) designates it as an input. Let \( y_{ij(\ell)} \) be the weight for each measure \( \ell \). One approach, presented by Cook and Zhu (2007), is to decide on the appropriate status of a flexible variable, is to view the problem from the perspective of the individual DMU. Specifically, adopt the position that for any given DMU the status should be that which maximizes that DMU’s efficiency score. Cook and Zhu (2007) establish the following mathematical programming model (4.4). Here each DMU is allowed to select the status of each variable that will credit it with the highest possible score. The authors then suggest taking
a majority rule position, giving each variable the status of inputs or outputs) preferred by the majority of the DMUs. An alternate model is given that views the problem from the perspective of the aggregate or composite of all the DMUs. The status of the flexible variable is that which maximizes the efficiency score of that composite DMU:

$$\max \sum \frac{x_{ij} \lambda_i}{\sum x_{ij} + \sum (1-\lambda_i) w_{ij}}$$

s.t. $$\sum x_{ij} \lambda_i + \sum (1-\lambda_i) w_{ij} \leq 1 \quad j = 1, 2, \ldots, n$$

$$d_i \in \{0, 1\}, \forall \ell \quad \mu_r, \nu_j, \gamma_i \geq 0, \quad \forall r, i, \ell.$$ (4.4)

### 6. Data variation

In the methodology developments discussed above, it is assumed that the data values are fixed and known. Significant literature has, however, been dedicated to situations wherein the data may exhibit “variation” or “uncertainty”; we briefly discuss various lines of research relating to such situations.

#### 6.1. Sensitivity analysis

This body of work addresses the question “if certain parameters/data within a model such as (2.1) and (2.2), or (2.3) are altered, how does this influence the efficiency status of DMUs?” Several directions have been taken here.

#### 6.1.1. Problem size issues

Various studies have examined the sensitivity of DMU efficiency to the addition of DMUs to or extraction of DMUs from the analysis. See Wilson (1995). A number of simulation studies (e.g. Banker et al., 1996) have related to the impact on the efficiency generated, for varying numbers of DMUs and of inputs and outputs.

#### 6.1.2. Direct data perturbations

Charnes and Neralic (1992) and Neralic (1997, 2004) addressed the subject of the impact on efficiency of perturbations to the values of inputs and outputs. We refer to this as involving direct data perturbation. That is, the research pertains to the derivation of ranges of variation in data over which matrix inversion in the simplex algorithm is unaffected.

#### 6.1.3. Indirect data perturbation – Radius of stability

An alternative to the above direct perturbation approaches, are what we should call indirect approaches. A number of studies have focused on the question of a maximum radius that will maintain the efficiency status. That is, “for a given DMU, what is the maximum allowable increase in outputs or decrease in inputs such that its efficiency status (efficient or inefficient) is unaltered?” The first work in this direction was initiated by Charnes et al. (1992). It examined the following problem relating to the additive model, and involving an inefficient DMU:

$$\max \delta$$

s.t. $$\sum \lambda_i x_{ij} + s_i^- = x_{ij} - \delta d_i^+, \quad i = 1, \ldots, m$$

$$\sum \lambda_j y_{rj} - s_j^+ = y_{rj} + \delta d_j^+, \quad r = 1, \ldots, s$$

$$\sum \lambda_j = 1,$$ (5.1)

$$d_i^-, d_j^+ = 1 \text{ or } 0, \text{ depending on whether a dimension is to be included or excluded from the perturbation } \delta.$$ This model looks at the maximum improvement in the status of an inefficient DMU (increase in outputs/decrease in inputs) before it is rendered efficient. For the case of an efficient DMU, Charnes and al examine the model:

$$\min \delta$$

s.t. $$\sum \lambda_i x_{ij} + s_i^- = x_{ij} - \delta d_i^+, \quad i = 1, \ldots, m$$

$$\sum \lambda_j y_{rj} - s_j^+ = y_{rj} + \delta d_j^+, \quad r = 1, \ldots, s$$

$$\sum \lambda_j = 1.$$ (5.2)

Here, we remove that DMU under evaluation, from the convexity considerations, much in the spirit of super-efficiency (Anderson and Petersen 1993).

These problems have been revisited by various authors including Cooper et al. (2001), Seiford and Zhu (1998a, 1999a), and others.

#### 6.1.4. Super-efficiency

An important problem in the DEA literature is that of ranking those DMUs deemed efficient by the DEA model, all of which have a score of unity. One approach to the ranking problem is that provided by the super efficiency model of Andersen and Petersen (1993), as mentioned above. See also Banker et al. (1989). The super-efficiency model involves executing the standard DEA models (CRS or VRS), but under the assumption that the DMU being evaluated is excluded from the reference set. In the input-oriented case, the model provides a measure of the proportional increase in the inputs for a DMU that could take place without destroying the “efficient” status of that DMU relative to the frontier created by the remaining DMUs.

The super-efficiency score can also be thought of as a measure of stability. That is, if input data for instance, is subject to error or change over time, the super-efficiency score provides a means of evaluating the extent to which such changes could occur without violating that DMU’s status as an efficient unit. Hence, the score yields a measure of stability.

In addition to being a tool for ranking, the super-efficiency concept has been used in other situations; for example, two-person ratio efficiency games (Rousseau and Semple, 1995), and acceptance decision rules (Seiford and Zhu, 1998b), among others.
It is well known that under certain conditions, the super-efficiency DEA model may not have feasible solutions for efficient DMUs (see, e.g., Zhu, 1996; Seiford and Zhu, 1998a, b; Dulá and Hickman, 1997; Seiford and Zhu, 1999a,b). As shown in Seiford and Zhu (1999a,b), infeasibility must occur in the case of the variable returns to scale (VRS) super-efficiency model. Although infeasibility implies a form of stability in DEA sensitivity analysis (Seiford and Zhu, 1998b), limited efforts have been made to provide numerical super-efficiency scores for those efficient DMUs for which feasible solutions are unavailable in the VRS super-efficiency model. Lovell and Rouse (2003) developed a standard DEA approach to the super-efficiency model by scaling up the inputs (scaling down the outputs) of a DMU under evaluation. As a result, a feasible solution can be found for efficient DMUs that do not have such (feasible) solutions in the standard VRS super-efficiency model. The super-efficiency scores for all efficient DMUs without feasible solutions are then equal to the user-defined scaling factor. Chen (2004, 2005) suggests using both the input- and output-oriented VRS super-efficiency models to quantify the super-efficiency when infeasibility occurs. However, Chen’s approach will fail if both the input- and output-oriented VRS super-efficiency models are infeasible.

Recently, Cook et al. (2008) posed an alternative approach to resolving the issue of infeasibility. Unlike the standard input-oriented and output-oriented super-efficiency models, each of which has a specific orientation (input or output), this model provides for the minimum movement in both directions needed to reach the frontier generated by the remaining DMUs. Viewed another way, in the case of infeasibility, the model derives the minimum change needed to project a data point, classified as an extremity, to a non-extreme position.

### 6.2. Data uncertainty and probability-based models

A number of researchers have concentrated on the problem of modeling technical efficiency when the data for the inputs and outputs are random variables. Thore (1987) and Land et al. (1992, 1994) looked at the application of chance constrained programming (CCP) to DEA. Cooper et al. (1996, 2004) demonstrated the use of CCP in the form of a satisficing model. Specifically, the CCR model (2.1) is replaced by the CCP model

\[
\text{max } P \left( \frac{\sum_{i} x_{ij} y_{ij}}{\sum_{i} x_{ij}} \geq \beta_{o} \right) \\
\text{s.t. } P \left( \frac{\sum_{j} y_{ij} x_{ij}}{\sum_{j} y_{ij}} \leq \beta_{j} \right) \geq 1 - a_{j}, \quad j = 1, \ldots, n \\
u_{r}, v_{i} \geq 0 \quad \forall r, i.
\]

(5.3)

Here the random variables \(\tilde{y}_{ij}, \tilde{x}_{ij}\) are assumed to have known probability distributions, and \(a_{j}\) is a scalar in the unit range [0,1] that specifies the allowable likelihood of failing to meet the constraints. The value \(\beta_{o}\) is referred to as an “aspiration level,” and specifies the desired efficiency level for \(\text{DMU}_{o}\).

### 6.3. Time series data – Window analysis

Time series data represent an important format in which “data variability” occurs. Specifically, in many applications, data for a DMU are available at different points in time, for example, in each of a set of quarters over several years. While one can perform static DEA analyses on the data for each quarter, and then apply standard regression concepts to study efficiency changes, such an approach often proves rather unsatisfactory, generally failing to capture important interactions from period to period. Window analysis as introduced by Charnes et al. (1985a) is a model structure that tries to bring a more robust treatment to efficiency changes in a time series sense. The idea is to choose a “window” of \(k\) observations for each DMU (say \(k = 4\) quarters), and treat these as if they represented \(k\) different DMUs. Hence, in any analysis, a total of \(n \times k\) “DMUs” are evaluated; \(k\) different scores for each DMU are then created. One then moves the window by one period (e.g. instead of quarters \(Q_{1}\) to \(Q_{4}\), one uses \(Q_{2}\) to \(Q_{5}\)) and repeats the analysis.

Window analysis allows the analyst to observe both the stability of a DMU for any point in time across different data sets, as well as trends across the \(k\) observations for each DMU, within the same data set. Cooper et al. (2006) discuss a number of weaknesses in conventional window analysis, one of which is the fact that beginning and ending periods are not tested as frequently as is the case for other periods. Sueyoshi (1992) has attempted to remedy this situation by the use of a “round robin” approach. This proceeds by first looking at each DMU in one period, then two, \(\ldots\), etc., up to \(k\) periods. This gives a more complete picture of stability and trends, but with the disadvantage of becoming computationally burdensome as the number of combinations grows exponentially.

### 6.4. Time series data – The Malmquist index

The Malmquist index was first suggested by Malmquist (1953) as a quantity for use in the analysis of consumption of inputs. Färe et al. (1994) developed a DEA-based Malmquist productivity index which measures productivity change over time. The index can be decomposed into two components, with one measuring the change in the technology frontier and the other the change in technical efficiency.

To describe the method, let \(x_{ij}^{t}, y_{ij}^{t}\) denote the input and output levels for a DMU, at any given point in time \(t\). The Malmquist index calculation requires two single period and two mixed period measures. The two single period measures can be obtained by using the CRS DEA model. Thus, for period \(t\) we solve following CRS DEA model which calculates the efficiency in time period \(t\), as displayed in (5.4)
the characterization of the production function by way of classical statistical inference methodology. Two lines of research have emerged around this issue: stochastic frontier analysis, and a DEA approach. Stochastic frontier regression dates back to Farrell (1957); this was subsequently extended by Aigner and Chu (1968) with their corrected ordinary least squares model. Later this approach was presented in a more formalized statistical format by Aigner et al. (1977), and has been labeled the “composed error” approach.

More recently Banker and Maindiratta (1992), Banker (1993) and Banker and Natarasan (2004) approached this issue from a DEA perspective. They show that DEA provides a consistent estimator of arbitrary monotone and concave production functions when the (one-sided) deviations from such a production function are regarded as stochastic variations in technical inefficiency. Convergence is slow, however, since, as is shown by Korostolev et al. (1995), the DEA likelihood estimator in the single output – m input case converges at the rate \( n^{-2/(1+m)} \) and no other estimator can converge at a faster rate.

The above approaches treat only the single output – multiple input case. Simar and Wilson (1998) turn to “bootstrap methods” which enable them to deal with the case of multiple outputs and inputs. In this manner, the sensitivity of \( \theta' \), the efficiency score obtained from the BCC model, can be tested by repeatedly sampling from the original samples. A sampling distribution of \( \theta' \) values is then obtained from which confidence intervals may be derived and statistical tests of significance developed.

In some respects, the output-oriented VRS model is a non-parametric version of the ordinary least squares model of Aigner and Chu (1968). This was alluded to in Banker and Maindiratta (1992) and Kuosmanen (2006).

For a thorough coverage of stochastic frontier analysis, and other approaches to efficiency evaluation, see Coelli et al. (1998), Kumbhakar and Lovell (2000).

7. Conclusions

This paper has attempted to provide a brief sketch of some of the important areas of research in DEA that have emerged over the past three decades. The focus here is on those topics that, in the authors’ estimation, have attracted the most attention. At the same time it is acknowledged that, due to limited space, and possibly to ignorance on our part, many important works in DEA may not have been highlighted. Some of these topics include the modeling of integer variables, issues of congestion, handling missing data, allocation of fixed inputs across DMUs, resource constrained DEA, analysis of composite DMUs, directional derivatives, and the connections relating to DEA and general multiple criteria decision models. In a number of these situations the topic has received some, but not significant attention, and may be a direction for future work.

Acknowledgment

The authors wish to thank Professor Robert Dyson, Editor, European Journal of Operational Research, for
his constructive comments on an earlier version of this article.

References