Chapter 1

Introduction

1.1 Background and Identification of the Problem

With increase in competition, organizations are under ever increasing pressure to increase their levels of operational efficiency. Conventional and traditional performance measurement systems generally focus on efficiency of an organization in absolute terms. This introduces a degree of arbitrariness in the resultant measure of efficiency. Moreover, traditional measures of efficiency have limited capacity of factoring in all the input and output measures that are understood to be contributors to the specific efficiency measure. Common methods of performance measurement use regression or stochastic frontier analysis. A constraining factor in the application of these techniques is the multiple inputs and outputs that often need to be considered.

Data Envelopment Analysis; a non-parametric, LP technique used in the estimation of production functions and has been used extensively to estimate measures of technical efficiency in a range of industries comes to the rescue in this scenario, by computing a measure of efficiency that is not absolute, but “relative” to other competing organizations.
Thus in DEA, an organization's efficiency is measured relative to the “best practices” in competing organizations. Further DEA adopts a method that can utilize all available, relevant inputs and outputs.

While “effectiveness” measures the extent to which organizations meet their objectives, “efficiency” is a measure of success with which an organization utilizes its resources (inputs) to meet its quantified objectives (outputs). A measure of efficiency could be either in technical terms, or in terms of allocation of resources or normated in terms of cost. It is also natural that the various inputs and outputs could have different units of measurement.

Organizations operating in the same domain, typically use an identical set of inputs that could include labor, capital, land, fuel and materials, to produce one or more outputs. For an organization that may not be using its inputs in a technically efficient manner, it might be possible to increase the quantities of outputs at the existing level of inputs, or to produce the existing level of outputs by utilizing reduced quantities of inputs. DEA measures efficiency by computing a measure of possible reduction in inputs or increase in outputs that is possible, given the competitive scenario, that is competitor data, that is available with the analyst.

A problem that can occur with DEA benchmarking is that the efficiency measure would be uncharacteristically high if there are few organizations that are being benchmarked relative to the number of explanatory variables.

Moffatt Associates (2006) mention another potential drawback with the use of benchmarking techniques that lies in the interpretation of the analysis results. Inefficiency is measured by the gap between the “efficiency frontier” and an organization. This is identical to either unnecessary costs or unreasonable output shortfall that is being caused by operational inefficiency. But inefficiency could also be due to conditions in
which the company operates.

In addition, the residual in terms of excess inputs or shortfall in outputs as compared to the DEA benchmark represents costs that the model fails to explain. This residual might be due to factors that have not been considered in the model. It must be remembered that benchmarking models cannot contain more than a few variables.

1.2 Theoretical Developments

This chapter discusses the primary tools used in this research, namely DEA and ANN. This is followed by a review of literature that has motivated and influenced this research.

1.2.1 Data Envelopment Analysis

Data envelopment analysis provides a means of measuring relative levels of efficiency among a group of organizations operating in a similar product or service domain. DEA computes this measure using Linear Programming by calculating the efficiency of an organization relative to the “best practice” observed within that group.

In order to set the perspective for DEA, it is necessary to look at various different efficiency concepts. By far the most common efficiency concept, technical efficiency, seeks to measure the conversion of inputs (such as the services of employees and machines) into outputs relative to the observed best practice. An organization operating at best practice is said to be 100% technically efficient. If an organization is operating below best practice levels, then the organization’s technical efficiency is expressed as a percentage of best practice. Technical efficiency is affected by operational practices and the scale of operations and is independent of prices and costs. Allocative efficiency
Chapter 1. Introduction

seeks to evaluate whether, for a certain given output level and a set of input prices, inputs are chosen to such that they minimize production cost, under assumptions of full technical efficiency. An organization operating at best practice in engineering terms could be allocatively inefficient if it is not using inputs in proportions that minimize costs. Cost efficiency refers to the combination of technical and allocative efficiency. An organization is cost efficient if it is both technically and allocatively efficient. Cost efficiency is calculated as the product of the technical and allocative efficiency scores (expressed as a percentage), so an organization can only achieve 100% cost efficiency if it has achieved 100% in both technical and allocative efficiency.

These concepts are depicted graphically, as in Figure 1.1 which plots different combinations of two inputs, labour and capital, required to produce a given quantity of output. An isoquant or ‘efficiency frontier” is smooth curve connecting the various combinations of minimum amounts of the two inputs required to produce the output. This represents the theoretical best engineering practice. An organization that is producing at any point on the isoquant is said to be technically efficient. The straight line denoted as the budget line plots combinations of the two inputs that have the same cost. The slope of the budget line is given by the negative of the ratio of the capital price to the labor price. Budget lines closer to the origin represent a lower total cost. Thus, the cost of producing a given output quantity is minimized at the point where the budget
lined is tangent to the isoquant. At this point both technical and allocative efficiencies are attained. The point of operation marked A would be technically inefficient because more inputs are used than are needed to produce the level of output designated by the isoquant. Point B is technically efficient but not cost efficient because the same level of output could be produced at less cost at point C. Thus, if an organization moved from point A to point C its cost efficiency would increase by \((OA - OA')/OA\). This would consist of an improvement in technical efficiency measured by the distance \((OA - OA')/OA\) and an allocative efficiency improvement measured by the distance \((OA' - OA'')/OA'\). Technical efficiency is usually measured by checking whether inputs need to be reduced in equal proportions to reach the frontier. This is known as a “radial contraction” of inputs because the point of operation moves along the line from the origin to where the organization is now.

1.2.1.1 Input-oriented Measures

This concept was illustrated by Farrell (1957) through an example of firms that use two inputs \((x_1 \text{ and } x_2)\) and produce one output \((y)\). This is illustrated in Figure 1.2 under the assumption of constant returns to scale.

The isoquant for efficient firms is represented by the curve \(SS'\) in Figure 1.2. The isoquant for full efficiency enables the measurement of technical efficiency. A firm at the point P uses the projected quantities in the \(x_1/y\) axis and \(x_2/y\) axis to produce unit

\[ \text{Figure 1.2: Technical and Allocative Efficiencies} \]
output. Its technical inefficiency
is the radial distance from the efficient frontier or isoquant and is given by the line
segment $QP$. $QP$ is the amount by which it should be possible to reduce both the inputs
– if we benchmark against the best-practice isoquant – without any reduction in the
output. In percentage terms the ratio $\frac{QP}{OP}$ represents the measure of inefficiency of the
firm at point $P$. From this, the technical efficiency of the firm is given by

$$TE = \frac{QP}{OP} = 1 - \frac{QP}{OP}; 0 \leq TE \leq 1$$

This indicates the extent of technical inefficiency of the firm, with 1 indicating full
technical efficiency. In the Figure 1.2, the point $Q$ which lies on the efficiency frontier is
technically efficient. The allocative efficiency could also be calculated, $AE = \frac{OR}{OQ}$, if
the ratio of input prices, represented by the slope of the line $AA'$ is known. Here $RQ$
represents the reduction in cost of production if instead of the production happening
at point $Q$, it were to occur at the allocatively efficient point $Q'$. The total economic
efficiency (EE) is defined to be the ratio $EE = \frac{OR}{OP}$ where the distance $RP$ can also
be interpreted in terms of a cost reduction. Note that the product of technical and
allocative efficiency provides the overall economic efficiency $TE \times AE = \left(\frac{OQ}{OP}\right) \times
\left(\frac{OR}{OQ}\right) = \left(\frac{OR}{OP}\right) = EE$. All three measures are bounded by zero and one; i.e.
$0 \leq TE, AE, EE \leq 1$. 

Computation of these measures of efficiency presupposes that the production function of the efficient firm is known. This is not the case in practice, where the sample data is used to construct the efficient isoquant. Farrell (1957) suggested using a non-parametric piecewise-linear convex isoquant (refer to Figure 1.3), or a parametric function be fitted to the data such that no observed point should lie on the origin side of the partition thus obtained.

1.2.1.2 Output-oriented Measures

The input-oriented technical efficiency measure addresses the question: “By how much can input quantities be proportionally reduced without changing the output quantities produced?” One could alternatively ask the question, “By how much can output quantities be proportionally expanded without altering the input quantities used?” This is an output-oriented measure as opposed to the input-oriented measure discussed above. The difference between the output- and input-oriented measures can be illustrated using an example involving one input and one output.

This is depicted in Figure 1.4(a) where we have decreasing returns to scale technology represented by $f(x)$, and an inefficient firm operating at the point P. The Farrell input-oriented measure of TE would be equal to the ratio $AB/AP$, while the output-oriented measure of TE would be $CP/CD$. The output- and input-oriented measures will only provide equivalent measures of technical efficiency when constant returns to scale exist.
but will be unequal when increasing or decreasing returns to scale are present Fare and
Lovell (1978). The constant returns to scale case is depicted in Figure 1.4(b) where we
observe that $AB/AP = CP/CD$, for any inefficient point P we care to choose. One can
consider output-oriented measures further by considering the case where production
involves two outputs ($y_i$ and $y$) and a single input ($x_i$). Again, if we assume constant
returns to scale, we can represent the technology by a unit production possibility curve
in two dimensions.

This example is depicted in Figure 1.5 where the line $ZZ'$ is the unit production possibility curve and the point A corres-
sponds to an inefficient firm. The inefficient point, A, lies below the curve in this case because $ZZ'$ rep-
resents the upper bound of pro-
duction possibilities.

The Farrell output-oriented efficiency measures would be defined as follows. In
Figure 1.5, the distance $AB$ represents technical inefficiency. That is, the amount by
which outputs could be increased without requiring extra inputs. Hence a measure
Chapter 1. Introduction

of output-oriented technical efficiency is the ratio $TE_0 = OA/OB$. If we have price information then we can draw the isorevenue line $DD'$, and define the allocative efficiency to be $AE_0 = OB/OC$ which has a revenue increasing interpretation (similar to the cost reducing interpretation of allocative inefficiency in the input-oriented case). Furthermore, one can define overall economic efficiency as the product of these two measures $EE_0 = (OA/OC) = (OA/OB) \times (OB/OC) = TE_0 \times AE_0$. Again, all of these three measures are bounded by zero and one. i.e. $0 \leq TE_0, AE_0, EE_0 \leq 1$.

1.2.1.3 Advantages and Limitations of DEA

The main advantage of DEA is that multiple inputs and outputs can be used for computing technical efficiency of a DMUs. For organizations that are observed to be inefficient, DEA identifies “peers” that could serve as potential role models, for improving its operations. However, a number of assumptions that are involved in DEA need to be considered in interpreting the results of a DEA study. DEA’s main limitations include the following:

- DEA results are sensitive to measurement error, and computes efficiencies that are relative to the observed best practices within the sample. Thus, it is meaningless to compare efficiency scores from different studies.

- DEA efficiency scores are sensitive to the specification of inputs and outputs, as well as the size of the sample used.

- Inspite of these limitations, DEA is a useful, and in many cases the only tool that can be used to measure the relative efficiencies of service providers or manufacturing units. This is a necessary precursor to understanding performance and coming up with means of improvement.
Chapter 1. Introduction

1.2.1.4 Data Envelopment Analysis

Seiford (1996) writes that in its present form, DEA was described in Charnes et al. (1978). Here, Charnes et al. (1978) establishes an efficiency frontier that is formed by a set of DMUs that exhibit “best practices” among the sampled DMUs. The process described then assigns efficiency levels to other DMUs based on their radial distances to the efficiency frontier. This method transforms multiple selected inputs and outputs into one index of efficiency. A wide range of variations in the computed efficiency index results from application of this basic idea. Different requirements for efficiency measurement are available; some of these are the constant returns to scale (CRS) model, the variable-returns-to-scale (VRS) model, the additive model, the slacks-based measures and the free disposal hull (FDH) model, etc. DEA has also been applied to various industrial and non-industrial contexts, such as banking, education, hospital, etc.

Surveyed DEA literature can be broadly categorized into three types – bibliography listing, qualitative, and quantitative. In Seiford (1997) and Gattoufi et al., 2004 Seiford (1997) and Gattoufi et al. (2004) provide extensive bibliography listings with an extensive list of DEA literature. Qualitative survey includes Seiford and Thrall (1990); Seiford (1997) in Seiford and Thrall (1990) and Seiford (1997). Cook and Seiford (2009) also provides a survey of DEA literature in Cook and Seiford (2009). While Seiford and Thrall (1990) in Seiford and Thrall (1990) review early-stage DEA development, Seiford (1996) traces the evolution of DEA from 1978 to 1995, describing the major achievements at each milestone. From a purely theoretical perspective, Cooper et al. (2007b) reviews some DEA and resultant efficiency measures. A comprehensive review on the methodological developments that have occurred in DEA since 1978 is documented by Cook and Seiford (2009). This encompasses most of the important DEA extensions.
Chapter 1. Introduction

such as generic DEA models, multilevel models, multiplier restrictions, considerations on the status of variables, and data variation. This analysis further notes two distinct phases in the development of DEA – pre and post 2000. The period after 2000 saw DEA developing at a speed that was faster than ever before, marked by the availability of DEA software - programs and packages. Some of these tools include DEAP by Tim Coelli, Frontier Analyst by Banxia Software, DEA Excel Solver by Joe Zhu, FEAR, a software package for Frontier Efficiency Analysis with R, developed by Paul Wilson, Bogetoft and Otto (2012), nonparaeff etc. These tools have made efficiency calculation easy and thereby removed a big entry-level hurdle for the field of DEA.

Recently there has been a spate of DEA applications used in performance evaluation of different types of entities such as universities, hospitals, business firms, countries, cities and regions in many countries. Since DEA requires very few strict assumptions, this technique has opened up possibilities for use in benchmarking in situations that have been hitherto resistant to analysis because of the complexity of relations between multiple inputs and outputs in the DMUs.

As pointed out in Cooper et al. (2007a), DEA has also been used to supply new insights into activities that have previously been evaluated by other methods. For instance, studies of benchmarking practices with DEA have identified numerous sources of inefficiency in some of the most profitable firms - firms that had served as benchmarks by reference to this (profitability) criterion – and this has provided a vehicle for identifying better benchmarks in many applied studies. Because of these possibilities, DEA studies of the efficiency of different legal organization forms such as ”stock” vs. "mutual" insurance companies have shown that previous studies have fallen short in their attempts to evaluate the potentials of these different forms of organizations. Similarly, a use of DEA has suggested reconsideration of previous studies of the efficiency with
which pre- and post-merger activities have been conducted in banks that were studied by DEA.

In their originating study, Charnes et al. (1978) described DEA as a ‘mathematical programming model applied to observational data [that] provides a new way of obtaining empirical estimates of relations - such as the production functions and/or efficient production possibility surfaces – that are cornerstones of modern economics’. Formally, DEA is a methodology directed to frontiers rather than central tendencies.

**Efficiency – Extended Pareto-Koopmans Definition:**

Full (100%) efficiency is attained by any DMU if and only if none of its inputs or outputs can be improved without worsening some of its other inputs or outputs.

In most management or social science applications the theoretically possible levels of efficiency will not be known. The preceding definition is therefore replaced by emphasizing its uses with only the information that is empirically available as in the following definition.

**Relative Efficiency Definition:**

A DMU is to be said to be “efficient” on the basis of data from sample DMUs if and only if, on the evidence of performances of other DMUs, none of its inputs or outputs can be improved without worsening some of its other inputs or outputs.

Literature on DEA today encompasses a variety of related approaches to evaluating performance that are rooted in DEA. The original CCR model has been extended and have provided deeper insights to the “multiplier side” from the dual formulation as well as the “envelopment side” from the primal formulation. Properties such as
Chapter 1. Introduction

non-concavity, piece-wise linearity, non-discretionary inputs, ordinal measurements, categorical variables, isotonicity, the Cobb-Douglas loglinear forms can be treated using DEA. relationships can also be treated through DEA.

We consider the input oriented DEA envelopment form. In this we have $n$ DMUs, each using $r$ inputs, usually denoted as the $X$ matrix, to produce $s$ outputs usually denoted as the $Y$ matrix.

1.2.1.5 The Envelopment Form

The LP formulation of the Envelopment form is as follows. The solution to this linear programming formulation produces a vector $\lambda$ where the non-zero elements correspond to the DMUs that form the efficient frontier envelop for DMU $X_0$. Let the peer DMUs for DMU $X_0$ be $\lambda = (\lambda_1, \lambda_2, \ldots, \lambda_k)$ where, in general, $k \ll n$, $n$ being the total number of DMUs. Let $X_{k,r+s}$ be the matrix where each row is the input, output vector for one of the efficient (peer) DMUs.

Then the input-output vector for the virtual benchmark DMU for $X_0$ is given by $\textbf{DMU}_{v0} = \lambda X$. This benchmark is based on quantities of input consumed and quantities of output produced by the DMUs. In the input oriented DEA, the benchmark consumes “not more than $\theta$ times any of the inputs while producing “at least” the amount of the
output produced by the evaluating DMU - $\theta$ minimized. Smaller values of $\theta$ indicate possibility of achieving identical output using lesser quantities of input, while a $\theta$ value close to 1 indicates near efficiency. $\theta \in [1, \infty)$ and the efficiency of the evaluating DMU in the input-oriented DEA is $1/\theta$.

1.2.1.6 The Multiplier Form

The multiplier form of DEA (corresponding primal) presents us with a vector of weights $(\mu_1, \mu_2, \cdots, \mu_r, \nu_1, \nu_2, \cdots, \nu_s)$ for each evaluating DMU, as also for DMU $X_0$. The LP formulation for the multiplier form is as follows. Each input-output weight vector

$$\max_{\mu, \nu} \mu Y_0$$

such that

$$\mu Y \leq \nu X$$

$$\nu X_0 = 1$$

$$\mu, \nu \geq 0$$

Figure 1.7: DEA: Multiplier Form

represents the preferred weights that a DMU assigns in the course of its benchmarking; such that the use of these weights to derive the values of inputs and outputs shows the evaluating DMU in the best possible light.

1.2.1.7 Extensions to the CCR Model

There are several application situations that can be analyzed using DEA. Some of these extensions that emphasise the adaptability of DEA are mentioned below.

The DEA extensions listed below allow the treatment of nondiscretionary as well as categorical variables. Extensions also make it possible to incorporate additional managerial information as well as value judgment. There are extension models that facilitate
Chapter 1. Introduction

the investigation of efficiency changes over time as well as to measure congestion.

1. Non-discretionary variables
2. Categorical variables
3. Incorporating a-priori knowledge
4. Window analysis

1.2.1.7.1 Nondiscretionary Inputs and Outputs

Applicability of DEA is driven by a general assumption that all the relevant inputs and outputs can be controlled by the DMU management. Hence the failure of a DMU to use minimum inputs to produce maximum outputs reduces the efficiency score. This ignores situations when there are exogenously fixed (non-discretionary) inputs and outputs that are beyond the DMU management’s control.

This model is implemented using the following formulation due to Banker and Morey (1986b) which recognizes that the degree to which a DMU may reduce a non-discretionary input may not be within the control or discretion of the DMU. This thus ceases to be an option in the corresponding DEA formulation.

Here, the inputs and outputs are divided into two subsets of discretionary \( (D_+) \) variables and non-discretionary \( (D_-) \) variables, where \( I_{D_+}, O_{D_+}, I_{D_-} \) and \( O_{D_-} \) refer to the discretionary input and output variables, and non-discretionary input and output variables respectively, with

\[
I = \{X_1, X_2, \cdots X_m\} = I_{D_+} \cup I_{D_-} \quad \text{with} \quad I_{D_+} \cap I_{D_-} = \phi
\]

\[
O = \{Y_1, Y_2, \cdots Y_s\} = O_{D_+} \cup O_{D_-} \quad \text{with} \quad O_{D_+} \cap O_{D_-} = \phi
\]

1.2.1.7.2 Categorical Inputs and Outputs

If one or more of the inputs and outputs are categorical the following extension of
the basic CCR model due to Banker and Morey (1986a) is used.

Let us assume that an input variable can assume \( L \) levels \((1, 2, \cdots L)\). These \( L \) levels divide the DMUs into \( L \) categories. Let us suppose that the set of \( n \) DMUs = \( \bigcup_{i=1}^{L} K_i \), where \( K_f = \{ j | j \in K, \text{and input value is } f \} \) and \( K_i \cup K_j = \emptyset, i \neq j \). Here a DMU is evaluated with respect to the DMUs in its own group and all preceding categories. Let \( DMU_0 \in K_f \). The following is the modified DEA formulation for categorical inputs.

\[
\begin{align*}
\text{Minimize} & \quad \theta \\
\text{such that} & \quad \sum_{j \in \bigcup_{f=1}^{K} K_f} x_{ij} \lambda_j + s_i^- = \theta x_{i0} \quad i = 1, 2, \cdots m \\
& \quad \sum_{j \in \bigcup_{f=1}^{K} K_f} y_{rj} \lambda_j + s_r^+ = y_{r0} \quad r = 1, 2, \cdots s \\
& \quad \lambda_j \geq 0 \quad j = 1, 2, \cdots n
\end{align*}
\]

The above specification evaluates all DMUs \( I \in D_1 \) with respect to the units in \( K_1 \), all DMUs \( I \in K_2 \) with respect to the units in \( K_1 \cup K_2 \), \cdots, and all DMUs \( I \in K_C \) with respect to the units in \( \bigcup_{f=1}^{K_C} K_f \), etc. This formulation assumes that there is a natural hierarchy of the categories. Each DMU is compared only with DMUs in its own and more disadvantaged categories, i.e. those operating under the same or worse conditions.

### 1.2.1.7.3 Incorporating a-priori knowledge

An important DEA extension restricts the allowable range for the DEA multipliers. In the CCR model, multipliers are only constrained to be positive. This flexibility sometimes gives rise to undesirable consequences, at times allowing a DMU to appear to be efficient whereas the same might be difficult to justify. In an unrestricted CCR scenario, the DEA model can assign unreasonably high or low values to some of the multipliers to maximize the efficiency rating for the evaluating DMU.
Chapter 1. Introduction

It is in such situations that restricting the range for multipliers can make the DEA results more realistic. There are many situations in which it has been seen beneficial to impose weight restrictions.

Dyson and Thanassoulis (1988); Roll et al. (1991) propose techniques that impose bounds on individual multipliers, while Thompson et al. (1986) imposes bounds on ratios of multipliers. Multiplier inequalities are used to impose weight restrictions in Wong and Beasley (1990) while Charnes et al. (1989) requires multipliers to belong to closed cones.

The standard multiplier DEA formulation, with the addition of additional inequality constraints is as follows:

Maximize \( z = \sum_{r=1}^{s} \mu_r y_{r0} \)

such that \( \sum_{r=1}^{s} \mu_r y_{rj} - \sum_{i=1}^{m} v_i x_{ij} \leq 0 \)
\( \sum_{i=1}^{m} v_i x_{i0} = 1 \)
\( \mu_r, v_i \geq 0 \)
\( \alpha_i \leq \frac{v_i}{v_{i0}} \leq \beta_i \quad i = 1, 2, \cdots m \)
\( \delta_r \leq \frac{\mu_r}{\mu_{r0}} \leq \gamma_r \quad r = 1, 2, \cdots s \)

Here, \( v_{i0} \) and \( \mu_{r0} \) represent multipliers in establishing the upper and lower bounds represented here by \( \alpha_i, \beta_i \) for the multipliers associated with inputs \( i = 1, \cdots, m \) and by \( \delta_r, \gamma_r \) for the multipliers associated with outputs \( r = 1, \cdots, s \) where \( \alpha_0 = \beta_0 = \delta_0 = \gamma_0 = 1 \).

The above constraints are called Assurance Region (AR) constraints as developed by Thompson et al. (1986) and defined more precisely in Thompson et al. (1990).
1.2.1.7.4 Window Analysis

Observations for DMUs are often available over multiple time periods. In such cases the interest focuses on changes in efficiency over time. In such a setting, it is possible to perform DEA over time by using a moving average analogue, where a DMU in each different period is treated as if it were a “different” DMU. Specifically, a DMU’s performance in a particular period is contrasted with its performance in other periods in addition to the performance of the other DMUs.

The window analysis technique that operationalizes the above procedure has been illustrated with the study of aircraft maintenance operations in Charnes et al. (1985). In this study, data were obtained for 14 (n=14) tactical fighter wings in the U.S. Air Force over seven (p=7) monthly periods. To perform the analysis using a three-month (w=3) window, one proceeds as follows.

Each DMU is represented as if it were a different DMU for each of the three successive months in the first window (M1, M2, M3). An analysis of the 42 (= nw=3×14) DMUs can then be performed. The window is then shifted one period by replacing M1 with M4, and an analysis is performed for these 42 DMUs on the second three-month set (M2, M3, M4). The process continues in this manner, shifting the window forward one period each time and concluding with the final (5th) analysis of 42 DMUs for the last three months (M5, M6, M7). (In general, one performs p-w+1 separate analyses, where each analysis examines nw DMUs).
1.2.1.7.5 Allocative and Overall Efficiency

The above models which consider “technical efficiency” do not require the use of prices or other “weights”. The “allocative efficiency” model extends the analysis to situations in which unit prices and unit costs are available; allowing incorporation of concepts of “allocative” and “overall” efficiency and relating them to “technical efficiency” in the manner first introduced by Farrell (1957). This model is implemented using the following formulation due to Cooper et al. (2007a).

Minimize \[ \sum_{i=1}^{m} c_{i0}x_i \]

such that \[ \sum_{j=1}^{n} x_{ij}\lambda_j \leq x_i, \quad i = 1, 2, \ldots m \]
\[ \sum_{j=1}^{n} y_{rj}\lambda_j \geq y_{r0}, \quad r = 1, 2, \ldots s \]
\[ L \leq \sum_{j=1}^{n} \lambda_j \leq U \]

resulting in

\[ 0 \leq \frac{\sum_{i=1}^{m} c_{i0}x_i^*}{\sum_{j=1}^{m} c_{i0}x_{i0}} \leq 1 \]

where \( x_i^* \) are optimal values obtained from above LP

and \( x_{i0} \) are observed inputs for \( DMU_0 \)

The objective is to minimize the total cost incurred by \( DMU_0 \) while satisfying the constraints. Here \( c_{i0} \) represent the unit costs of inputs for \( DMU_0 \) and are allowed to vary across the DMUs. The envelopment parameters, \( L \) and \( U \) constrain \( \sum_j \lambda_j \). Choosing \( L = U = 1 \) makes this a BCC model, while removing the \{L,U\} constraint, i.e. \( L=0, U=\infty \) makes this a CCR model.
Chapter 1. Introduction

Using the standard approach the measure of relative cost or relative efficiency is obtained by utilizing the ratio

\[ 0 \leq \frac{\sum_{i=1}^{m} c_{i0} x_i^*}{\sum_{i=1}^{m} c_{i0} x_{i0}} \leq 1 \]

where \( x_i^* \) are optimal values obtained from above LP formulation and \( x_{i0} \) are observed inputs for \( DMU_0 \)

1.2.1.7.6 Profit Efficiency

We now introduce another type of model called the “additive model” to evaluate technical inefficiency. First introduced in Charnes et al. (1985), this model has the form:

Maximize \( \sum_{r=1}^{s} s_r^+ + \sum_{i=1}^{m} s_i^- \)

such that \( y_{r0} = \sum_{j=1}^{n} y_{rj} \lambda_j - s_r^+ \quad r = 1, 2, \ldots s \)

\( x_{i0} = \sum_{j=1}^{n} x_{ij} \lambda_j - s_i^- \quad i = 1, 2, \ldots m \)

\( \sum_{j=1}^{n} \lambda_j = 1 \)

\( 0 \leq \lambda_j, s_r^+, s_i^- \quad \forall i, j, r \)

This model dispenses with the need for distinguishing between an “output” and an “input” orientation because the objective simultaneously maximizes outputs and minimizes inputs. This can be seen by utilizing the solution to introduce new variables
The slacks are all independent of each other. Hence an optimum is reached only when it is impossible to increase an output \( \hat{y}_{r0} \) or reduce an input \( \hat{x}_{i0} \) without decreasing some other output or increasing some other input.

### 1.2.2 Extensions to DEA

DEA applications have revealed some drawbacks. Important among them are the following.

1. When there are few DMUs and comparatively a large number of variables, DEA is unable to discriminate among efficient DMUs.
2. Sometimes the weighting scheme can be unrepresentative and not in consonance with the relative importance of the variables.
3. Sometimes DEA can result in more than one optimal solution corresponding to the weighting scheme of extreme efficient DMUs.

There are a few approaches that have been used to alleviate these shortcomings. These methodologies that extend DEA can be classified into two groups:

1. Methods incorporating decision-makers a-priori information in the model
   
   (a) weight restriction methods
   
   (b) preference structure method
   
   (c) value efficiency analysis
2. methods not requiring such a-priori information

   (a) super efficiency method
   (b) cross-evaluation method
   (c) multiple-objective linear programming (MOLP) method

1.2.2.1 Methods incorporating a-priori information

   The preferences of the decision-maker about the importance of the input-output variables can be incorporated in the DEA model.

   The Weight Restriction methods described below incorporate such value judgments in DEA.

1.2.2.1.1 Weight Restriction: Direct Weight Restriction

   Dyson and Thanassoulis (1988) developed this approach in which the restrictions are of the type:
   \[ \alpha_i \leq v_i \leq \beta_i \]  for input \( i \)
   \[ \alpha_r \leq u_r \leq \beta_r \]  for output \( r \)

   The restrictions impose numerical limits on the weights and ensure that none of the variables are either ignored or overestimated in the DEA analysis. These bounds would depend on the context of the problem and would be based on information provided by an expert. Hence, these bounds can only be established after analysing the weights from the original DEA problem. It is important to remember that direct weight restricted DEA models produce differing efficiency measures based on the model orientation.
1.2.2.1.2 \textbf{Weight Restriction: Cone Ratio}

Charnes et al. (1989, 1990) developed the Cone Ratio variant of weight-restricted models. This variation remedies, to some extent, the problems of overestimation of efficiencies when the weights selected are at variance with the DMU objectives. The Cone Ratio method determines weights that are more consistent with the objectives of the DMUs.

Let $V = A'\alpha$, where $A' = (a_1, a_2, \ldots, a_k) \in \mathbb{R}^{m \times k}$ and let $\alpha' = (\alpha_1, \ldots, \alpha_k)$ be the polyhedral cone, where $a_j$ are the direction vectors ($j = 1, \ldots, k$). In the same way, we can define a polyhedral cone for $u$ as $U = B'\beta$, where $B' = (b_1, \ldots, b_n) \in \mathbb{R}^{m \times n}$ and $\beta' = (\beta_1, \ldots, \beta_n)$.

Given the standard CCR model:

\[
\begin{align*}
\text{Maximize} & \quad u'y_0 \\
\text{such that} & \quad -v'X + u'Y \leq 0 \\
& \quad v'x_0 = 1 \\
& \quad v \in V \\
& \quad u \in U
\end{align*}
\]

the Cone Ratio model is formulated as follows in terms of $\alpha$ and $\beta$.

\[
\begin{align*}
\text{Maximize} & \quad \beta'(B'y_0) \\
\text{such that} & \quad -\alpha'(AX) + \beta'(BY) \leq 0 \\
& \quad \alpha'(Ax_0) = 1 \\
& \quad \alpha, \beta \geq 0
\end{align*}
\]

Thus the Cone Ratio model is in effect a CCR model on transformed DMU data and results in at least one efficient DMU. For interpretation, the DEA results must be
transformed back into the original form.

1.2.2.1.3 Weight Restriction: Assurance Region

The concept of Assurance Regions was developed by Thompson et al. (1986) by introducing homogeneous, linear restrictions. This approach works by iteratively refining the assurance region until the resultant efficiency levels appear to be satisfactory to the decision-maker. The values of the upper and lower bounds are dependent on the measurement scales of the inputs and outputs. It has also been shown that the Assurance Region is a special case of Cone Ratio models and like the Cone Ratio model, results in at least one efficient DMU.

1.2.2.1.4 Weight Restriction: Virtual input and output restriction

The previous weight-restriction methods do not take into consideration the input-output levels of the DMUs. The derived weights would thus be dependent on the input-output levels and would prevent establishing a relationship among the several weights. The Virtual input and output restriction method, on the other hand, considers restrictions on the “virtual variables”, i.e. the derived DMU weight times the level of the variable, instead of directly restricting the weights.

The use of virtual input and output restrictions has been explored by Wong and Beasley (1990), who restrict the proportion of virtual output $r$ (weight times variable value) to the total virtual output for any DMU. For $DMU_j$ and output $r$, this translates to:

$$\alpha_r \leq \frac{u_r y_{rj}}{\sum_{r=1}^{s} u_r y_{rj}} \leq \beta_r$$
Chapter 1. Introduction

where \( \sum_{r=1}^{s} u_{r} y_{rj} \) is the total virtual output of the DMU \( j \). Similar set of restrictions can be obtained for the virtual inputs.

The implementation of this kind of restriction is indirect and necessitates an earlier DEA run. To avoid this necessity, Wong and Beasley (1990) suggest the following modifications:

1. This restriction is added only to the DMU that is being evaluated. This means that each DMU is analyzed with two additional restrictions

2. This restriction is added to all DMUs; which means that for \( n \) DMUs, there are \( 2n \) additional constraints. This makes it computationally expensive.

3. This restriction is added to the proportion of the total virtual output that the “average” DMU gives to output \( r \). In this scenario, every DMU is analyzed with the following two additional restrictions:

\[
\phi_{r} \leq \frac{u_{r} \sum_{j=1}^{N} (y_{rj}/N)}{\sum_{r=1}^{s} u_{r} \left( \sum_{j=1}^{N} (y_{rj}/N) \right)} \leq \psi_{r}
\]

where \( \sum_{j=1}^{N} (y_{rj}/N) \) is the level of the \( r^{th} \) output of the “average DMU”.

However in each of the weight-restriction models the efficiency scores are sensitive to the DEA model orientation.

1.2.2.1.5 Preference Structure method

Zhu (1996) presented models using weights to introduce preference structure in DEA. In this scenario, the preference structure prescribed by the decision-maker drives
the targets for inefficient DMUs. This target, based on an equi-proportionate reduction in inputs or increment in outputs is more meaningful than that produced by classic DEA models, and hence provides a more authentic efficiency score.

Three models have been proposed by Zhu (1996), of which the following model combines the two orientations:

Maximize \[ \sum_{r=1}^{s} w^+_r \phi_r - \sum_{i=1}^{m} w^-_i \theta_i \]

such that \[ \sum_{j=1}^{n} \lambda_j y_{rj} - s^+_r = \phi_r y_{rj0} \quad r = 1, \ldots, s \]
\[ \sum_{j=1}^{n} \lambda_j x_{ij} - s^-_i = \theta_i x_{ij0} \quad i = 1, \ldots, m \]
\[ \phi_r, \theta_i \text{ free} \quad \forall r, i \]
\[ s^-_i, s^+_r \geq 0 \quad \forall i, r \]

where \[ \sum_{r=1}^{s} w^+_r - \sum_{i=1}^{m} w^-_i = 1 \]

In this model, the set of weights, \( w^+_r, w^-_i \) is the decision-maker’s preference structure. The objective function value gives the efficiency score for the DMU that is being analyzed.

1.2.2.1.6 Value efficiency analysis

This is a two-stage developed by Halme et al. (2000) to incorporate the decision-maker’s preferences into the DEA analysis. In the first stage the solution that the decision-maker prefers most is identified through a multiple objective model. In the second stage the frontier is determined as the point at which the decision-maker’s value function assumes its maximum.
Chapter 1. Introduction

According to Halme et al. (2000), this “most preferred solution” is computed using the multiple objective linear program:

\[
\text{Maximum } U\lambda = \begin{bmatrix} Y \\ -X \end{bmatrix} \lambda
\]

such that \( \lambda \in \Lambda = \{ \lambda | \lambda \in \mathbb{R}^n, A\lambda \leq b \} \)

This multiobjective model depends on a decision-maker to determine which solution is most appropriate and hence does not have any single solution. The model facilitates the decision-maker to identify the “most preferred solution” which would be either a virtual or an existing DMU on the efficient frontier with preferred input/output levels chosen by the decision-maker.

With the “most preferred solution” chosen, the following LP formulation is used to determine the value efficiencies for an individual DMU.

Maximize \( \sigma + \varepsilon (1's^+ + 1's^-) \)

such that \( Y\lambda - \sigma w^s - s^+ = g^y \)
\( X\lambda + \sigma w^x - s^- = g^x \)
\( \sum \lambda = 1 \)
\( s^+, s^- \geq 0 \)
\( \varepsilon > 0 \)
\( \lambda_j \geq 0, \text{ if } \lambda_j^* = 0, \quad j = 1, \cdots, n \)

where the MPS is \( y^* = Y\lambda^*, x^* = X\lambda^* \)

Further developments of this approach such as Halme and Korhonen (2000); Korhonen et al. (1998) incorporate weight-restrictions and situations where more than one
first-stage solution is desirable to the decision-maker.

After including preference information, a better discrimination among DMUs is obtained.

1.2.2.2 Methods not requiring a-priori information

Subjectvity is a major disadvantage in methods incorporating a-priori information.

1. The value judgements captured in the a-priori information may be biased or not in sync with reality

2. Decision-makers may not have a consensus about the a-priori information and this could delay and adversely affect the study

The following methods do not require a-priori information and thereby avoid or minimize the intervention of the expert but at the same time, increase discrimination in DEA.

1.2.2.2.1 Super-efficiency method

This method introduced by Andersen and Petersen (1993) compares the DMU being evaluated with linear combination of other DMUs excluding the DMU being evaluated. This approach affects the efficiency scores of extreme efficient DMUs which can obtain efficiency scores greater than one. This approach provides a ranking of efficient DMU by allowing scores in excess of one.
Chapter 1. Introduction

The CRS input oriented super-efficiency model is:

\[
\begin{align*}
\text{Minimize} & \quad \theta - \varepsilon (1's^+ + 1's^-) \\
\text{such that} & \quad \sum_{k=1, \neq j}^{n} \lambda_k X_k + s^- \quad \theta X_j \\
& \quad \sum_{k=1, \neq j}^{n} \lambda_k Y_k + s^+ \quad Y_j \\
& \quad \lambda_k, s^+, s^- \geq 0
\end{align*}
\]

This methodology allows discrimination among efficient DMUs and a ranking for all DMUs but it does not solve the problem of unrealistic weights.

1.2.2.2.2 Cross-evaluation

Another method developed by Sexton (1986) increases discrimination among efficient DMUs by using cross-evaluation. Unlike in classic DEA models, in “cross-evaluation” DEA is used in peer-evaluation instead of self-evaluation. In a peer-evaluation each DMU is evaluated according to the DEA computed weighting scheme of other DMUs and the “cross-efficiency” is the average of these efficiencies.

In the DEA models, the possibility of multiple optimal solutions can cause cross-efficiencies to vary since one scheme can be favourable to one DMU and unfavorable to another. This calls for the use of a secondary objective function. For the secondary objective function, there are two alternative formulations: the “aggressive formulation” in which the secondary objective, the “cross-efficiencies” of the other DMUs, are minimized, and the “benevolent formulation” in which the secondary objective function is maximized.

We define \( E_{ks} = \frac{\sum_{i} u_{ki} y_{si}}{\sum_{i} v_{kj} x_{sj}} \) as the cross-efficiency of DMUs using \( DMU_k' \)'s weighting scheme. Thus \( E_{kk} \) would be the efficiency score for \( DMU_k \) using its own weighting.
scheme, i.e., the efficiency score calculated by the CCR model. This efficiency would be called “standard efficiency”.

There are three forms of the “aggressive formulation”. In one approach, for $DMU_k$, we minimize the mean of the cross-efficiencies of others DMUs using the chosen weighting scheme that maximizes the objective function in the first step, resulting in the following second stage objective:

\[
\text{Minimize } (n - 1)A_k = \sum_{s \neq k} E_{ks} = \sum_{s \neq k} \sum_i u_{ki} y_{si} - \sum_j v_{kj} x_{sj}
\]

The objective function above is non-linear and is converted, for the purposes of using a LP model, using the standard substitution below. Converting the above non-linear objective function to a linear form, we have the following LP formulation.

Minimize $B_k = \sum_{s \neq k} \left( \sum_i u_{ki} y_{si} - \sum_j v_{kj} x_{sj} \right)$

\[
= \sum_i (u_{ki} \sum_{s \neq k} y_{si}) - \sum_j (v_{kj} \sum_{s \neq k} x_{sj})
\]

such that

\[
\sum_i u_{ki} y_{ki} - E_{kk} \sum_j v_{kj} x_{kj} = 0
\]

\[
\sum_i u_{ki} y_{si} - \sum_j v_{kj} x_{sj} \leq 0
\]

\[
\sum_j v_{kj} x_{kj} = 1
\]

$u_{ki}, v_{kj} \geq 0$

where $u_{ki}, v_{ki}$ are the decision variables
The Cross-efficiencies matrix table, Table 1.1 shows the cross-efficiencies thus calculated. Here $E_{kk}$ is the classic DEA efficiency while $E_{ks}$ is the efficiency score calculated using the weighting scheme obtained for $DMU_k$, and $e_k$ is the average cross-efficiency of $DMU_k$:

$$e_k = \frac{1}{n} \sum_{s} E_{sk} \text{ i.e., the average of the efficiencies in column } k \text{ of the } 1.1, \text{ allowing self-evaluation of } DMU_k.$$ 

Sometimes $e_k$ is calculated without considering the classic DEA efficiency:

$$e'_k = \frac{1}{n-1} \sum_{s \neq k} E_{sk}.$$ 

Besides using this measure as a complement or alternative to the standard efficiency in order to distinguish among efficient DMUs, one can also distinguish DMUs with the greatest difference between the standard efficiency and mean cross-efficiency, which can be done by:

$$M_k = \frac{(E_{sk} - e'_k)}{e'_k}.$$ 

Though this “cross-evaluation” method does not requiring a-priori information it is complex requiring the solution of a second model in order to calculate the average cross-efficiencies. Moreover an addition or exclusion of a DMU can alter the final cross-efficiencies. This is why this method is appropriate only in case it it is unlikely to have any changes in the set of DMUs.

1.2.2.2.3 Multiple objective approach

A multiple objective approach, called “Multiple Criteria Data Envelopment Analysis” – MCDEA, is proposed by Li and Reeves (1999). This approach focuses on solving two main problems:
Chapter 1. Introduction

- lack of discrimination
  which happens when the number of variables is large compared with the number of DMUs – resulting in far too many DMUs being identified as “efficient”
- inappropriate weighting schemes
  which happens when an “efficient” DMU assigns near-zero weights to some variables

The above two related problems often occur together. The large number of variables results in each DMU assigning weights to a few variables in order to attain efficiency – thus assigning zero weights to the other variables.

In order to address the above problems, Li and Reeves (1999) include two additional objective functions in the standard DEA model. The DMUs are thereby evaluated in a Multiobjective Linear Programming context.

The classic DEA CCR model for $DMU_0$ is:

Maximize $h_0 = \sum_{r=1}^{s} u_r y_{rj_0}$

such that $\sum_{r=1}^{s} u_r y_{rj} - \sum_{i=1}^{m} v_i x_{ij} \leq 0 \quad j = 1, \cdots, n$

$\sum_{i=1}^{m} v_i x_{ij_0} = 1$

$u_r, v_i \geq 0 \quad \forall r, i$

$DMU_0$ is efficient if and only if $h_0 = 1$.

Li and Reeves (1999) introduce two additional objective functions; the first of these seeks to minimize the maximum deviation for which the restriction included in the new formulation, $M - d_j \geq 0 \quad (j = 1, \cdots, n)$, makes $M$ the maximum deviation. The second additional objective seeks to maximize the deviation of all DMUs. The resulting problem is:
Chapter 1. Introduction

Maximize \[ h_0 = \sum_{r=1}^{s} u_r y_{rj_0} \]

Minimize \[ M \]

Minimize \[ \sum_{j=1}^{n} d_j \]

such that \[ \sum_{r=1}^{s} u_r y_{rj} - \sum_{i=1}^{m} v_i x_{ij} \leq 0 \quad j = 1, \cdots, n \]

\[ M - d_j \geq 0 \quad j = 1, 2, \cdots, n \]

\[ \sum_{i=1}^{m} v_i x_{ij_0} = 1 \]

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

While LP models result in a unique optimal solution as in classic DEA, this model provides a set of non-dominated (or efficient) solutions. From this set of efficient or non-dominated solutions solutions that optimize each of the three objectives can be found. Attaching some weight to the second and third objectives would reduce efficiencies of all DMUs in general and for “efficient” DMUs in particular.

The addition of the two additional objective functions results in increased discrimination using DEA.

Stability of efficiency scores of DMUs is indicated by the number of non-dominated solutions. The greater the number of non-dominated solutions, the higher is the sensitivity of efficiency scores to changes in criteria.

The total number of non-dominated solutions associated to a DMU generally reflects the stability of the efficiency scores of the DMU relative to changes in the efficiency criteria, that is, the greater the quantity of solutions, the more sensitive the efficiency scores of a DMU are to changes in the criteria.

The selection of one solution among a set of solutions can be seen in this case as a way of including preferences, even though this method does not per-se call for a-priori
information. This is different from the approach in Value Efficiency Analysis where a MPS is first identified in order to determine the efficient frontier and the final solution.

1.2.3 Neural Networks

1.2.3.1 Biological Neural Network

Artificial Neural Networks refer to computational models inspired by animals’ central nervous systems (in particular the brain) that are capable of machine learning and pattern recognition and usually presented as systems of interconnected “neurons” that can compute values from inputs by feeding information through the network. An ANN is composed of a large number of interconnected neurons. The architecture of the ANN has been inspired by the human brain. While an individual neuron can perform a simple task, when a network of connected neurons is capable of performing complex tasks, with great speed and accuracy.

As shown in Figure 1.8 a neuron consists of a cell body called the soma which contains the nucleus, dendrites and an axon. The dendrites receive signals from other neurons.

The axon transmits signals to neighbouring neurons. The synapse, which is the communicating unit between neurons, connects the axon of one neuron with the dendrites of a neighboring neuron. A synapse carries electrochemical signals from one neuron to another. When the signal exceeds the synapse threshold, the neuron fires and sends an electrochemical signal to a neighbouring

Figure 1.8: Biological Neuron
neuron. The basis of memory is realized through changing of the strength of synaptic connections.

1.2.3.2 Overview of Artificial Neural Networks

The ANN is modeled on the biological neural network. It is an interconnection of nodes which are analogous to neurons. There are fundamentally three critical components of a neural network:

- node character
determines how signals are processed by the node; the weight associated with each connection and the activation function

- network topology
determines how the nodes are organized in the various layers and the connected

- learning rules
determine how the weights are initialized and how they are adjusted to arrive at the equilibrium or final state

1.2.3.2.1 Node Character The model for a ANN node is depicted in Fig. 3.4. Every node receives inputs via weighted connections from other nodes. The weights are analogous to the strength of the synapse. A node activates or fires when the weighted sum of its inputs exceeds its activation threshold. On activation, the signal is passed through the transfer function to neighbouring nodes. A mathematical model for this process is:

\[ y = f(\sum_{i=0}^{n} w_ix_i - T) \]
where \( y \) is the node’s output, \( f \) is the transfer function, \( w_i \) is the weight of input \( x_i \), and \( T \) is the value of the activation threshold. The simplest transfer function is the step function:

\[
y = \begin{cases} 
0 & \text{if } \sum_{i=0}^{n} w_i x_i > T \\
1 & \text{if } \sum_{i=0}^{n} w_i x_i < T 
\end{cases}
\]

Another commonly used transfer function is the sigmoid function, Figure 1.9, which has the advantage of itself as well as its derivative being continuous.

The transfer function, in this case, is of the form

\[
\frac{1}{1 + e^{-\beta x}}
\]

In this section we provide an outline of ANN, which is drawn from Zou et al. (2008).

The feedforward network, also known as perceptron is one of the most used type of neural networks and has a basic architecture as in Figure 1.10. It consists of a set of neurons, also called processing units, which are arranged
Chapter 1. Introduction

in layers.

Between the input and the output layers are one or more “hidden” layers. The neurons are connected as shown with each connection representing a weight. In the Figure 1.10, there are four inputs, one “hidden layer” containing three neurons and an output neuron. There are 15 connections (weights) \(15 = 4 \times 3 + 3\).

The neurons are information-processing units that are fundamental to the operation of a neural network. Figure 1.11 shows the model of a neuron.

There are three basic processes of the neuron model:

1. each input signal, \(x_j\), is multiplied by the corresponding synaptic weight, \(w_j\)
2. summation of the weighted input signals, \(\sum_{j=1}^{p} x_j w_j\)
3. applying the activation function to the sum

The activation function defines the output of a neuron as a function of the activity level at its input. The sigmoid function is a common form of activation function used in ANNs. An common sigmoid function is the logistic function, defined as:

\[
f(x) = \frac{1}{1 + e^{-x}}
\]

A neural net is defined by the number of layers, the number of neurons in each layer, and the connection weights. The process of weight estimation is called “training”.

Figure 1.11: Model of a Neuron
1.2.3.3 Neural Network Training

A neural network learns a model from a training set presented to it. Each training data consists of an input–output pair: an input signal and the corresponding response of the neural network. The NN outputs of each training data are compared with the desired outputs. If \( y = (y_1, y_2, \ldots, y_p) \) represents the outputs (\( p \) is the number of neurons in the output layer), and \( d = (d_1, d_2, \ldots, d_p) \) is a vector of the desired response, the error for the \( k^{th} \) data is:

\[
e_k = \frac{1}{p} \sum_{j=1}^{p} (y_j - d_j)^2
\]

With \( n \) data examples, the total error is

\[
e = \frac{1}{n} \sum_{k=1}^{n} e_k
\]

The objective is to iteratively reduce the error by updating the weights.

1.2.3.4 Backpropagation

The basic steps of neural network training using the backpropagation (steepest descent) algorithm are:

1. initialize the network weights to small uniformly distributed random numbers
2. present the examples to the network and compute the outputs
3. compute the error
4. update the weights backward, i.e., starting from the output layer and passing
the layers to the input layer, using the delta rule (Figure 1.12): \( \Delta w_{ji} = -\eta \frac{\partial e}{\partial w_{ji}} \)

where \( \eta \) is a constant representing the learning rate.

### 1.2.3.5 Network Topology

Figure 1.13 shows the general architecture of an ANN. The nodes are organized into layers in linear arrays. There are three types of layers - input layers, output layers, and optional hidden layers. A network can have none or multiple hidden layers. The
design of a network topology involves determining the number of hidden layers, the number of nodes in each layer and the connections between the nodes in adjacent layers. The network topology may be initially set by intuition and optimized through multiple experimental iterations. There are instances of rational methods used in the design of a neural network, e.g. Chiu and So (2003) describe a genetic neural network to solve a class of problems using a genetic algorithm to select the input features for their neural network.

There are two types of connections between nodes that characterize a neural network as “feedforward” or “feedback”. A feedforward network comprises one-way connections with no loop back and requires the signal to travel in one direction only. Except during training of the neural network, there are no backward links. This type of network is static – one input is associated with one particular output. A widely used feedforward network is the “Perceptron”. A feedback network, on the other hand, comprises loop-back connections in which the output of nodes can loop back as inputs to previous layer nodes or nodes in the same layer. This type of a neural network is dynamic and for any input the state of a feedback network changes for many cycles, ultimately reaching an equilibrium. In a feedback network one input thus produces a series of outputs. Examples of widely used feedback networks are the Hopfield network and the Kohonen Self-Organizing Maps (SOM).

1.2.3.6 Learning in Neural Networks

The ANN needs to undergo a phase of “training” before it can be used. The weights that transform inputs to final outputs are adjusted to the optimal or desired values during this training. There are two main types of learning – supervised learning and unsupervised learning. In supervised learning, a neural network is provided examples of
inputs and corresponding outputs. These examples or “training set” typically represent
the underlying network model that is sought to be designed. During the course of
“training”, the network weights are iteratively adjusted so as to minimize the error, or
the difference between the network output and the supplied observed output for the
training set. It is important to have a representative training set, otherwise the trained
neural network is unlikely to be reliable or sufficiently general. Ultimately, when the
neural network produces the desired outputs for the “training set” inputs, the computed
weights are finalized and the network is operationalized.

In contrast to supervised learning networks that require to undergo an initial training
phase, unsupervised networks do not use target outputs from a training set. Instead the
unsupervised network “discovers” the underlying patterns or trend in the input data.

Two of the more common learning approaches for ANNs are “error correction” and
“nearest neighbour” methods.

The “error correction” method relies on the back propagation mechanism. Let $y_{k,n}$
be the $n^{th}$ step output of the $k^{th}$ output node and $y^*_k$ be the target output for the $k^{th}$
node. An error function can be defined as the difference between the node output and
target output:

$$ e_k = y_{k,n} - y^*_k $$

Let $\lambda$ be a positive constant that modulates the rate of weight adjustment. The
new weight for input $x_j$ is calculated as $w_{k,j,n+1} = w_{k,j,n} - \lambda e_k X_j$.

The weight vector is updated every step until the system converges. The rate of
learning, $\lambda$, has major impact on the system converging rate.

Another approach of ANN learning is based on the spatial similarity and is commonly
used for classification problems. The entire training set data are stored inside the network
memory. For any test sample \( x_t \), its nearest neighbour must have a distance that satisfies

\[
\min_{i=1}^{N} D(X_i - X_t),
\]

where \( D \) is a distance measurement function and \( N \) is the size of the training sample. The test sample assumes properties similar to its closest neighbor based on the predefined function. Though the nearest neighbour learning approach is simple, as the dimension of the problem increases, the error rate often becomes unacceptable.

### 1.2.3.7 Common Artificial Neural Network Models

#### 1.2.3.7.1 Perceptron

One of the earliest neural networks is the “perceptron” – a feedforward network developed by Rosenblatt (1962). A single hidden layer perceptron has severe limitations and can solve only linearly separable problems. This has been demonstrated using the classic XOR problem, which Minsky and Papert (1969) show to be unsuitable for simulation using a single-layer perceptron.

![Multilayer Perceptron](image)

**Figure 1.14: Multilayer Perceptron**

The multilayer perceptron (MLP), shown in Figure 1.14, is the most widely used
neural network and can be used to approximate any continuous functions. In the training of a multilayer perceptron a backpropagation algorithm is generally used after Rumelhart and McClelland (1986). The input is propagated through the network and used to calculate the output. Thereafter the “cost” represented by the error or the difference between the correct and the computed output is propagated back from output to input layer in order to adjust the network weights. Since this algorithm minimizes the cost function with a gradient descent method, it can be only be applied to networks with differentiable transfer functions and it is therefore often difficult to train.

1.2.3.7.2 Hopfield Network

The Hopfield net (Hopfield (1982)), shown in Figure 1.15 is a feedback neural network with only a single layer. A Hopfield net is fully connected, i.e. every node is connected with every other node except itself. For a given input, a Hopfield net is repeatedly updated until it converges to an equilibrium state, known as the “attractor state”. Any arbitrary input eventually leads the network to an attractor states. The classic “Travelling Salesman Problem” can be solved using the Hopfield net.
1.2.3.7.3 Kohonen Map

The Kohonen map, vide Kohonen (1989), is another feedback network. It has a single hidden layer with nodes typically arranged in a two-dimensional grid layout as shown in Figure 1.16. Each input node connects to every node in the hidden layer and each node in the hidden layer. The Kohonen map uses unsupervised learning with the winner-takes-all learning rule. The node with the highest response and its neighbouring nodes get their weights updated, making them more likely to respond to similar input patterns. Kohonen maps can be used for projection of high-dimensional data into two-dimensional space and data clustering.

1.3 Literature Review

This section is structured on the basis of the researcher’s study of existing work on DEA (section: DEA), ANN (section: Artificial Neural Network), and publications involving integration of various techniques, such as DEA, ANN, cluster analysis, Multi-objective Linear Programming; an area of MCDM, that is concerned with mathematical optimization involving more than one objective function to be optimized simultaneously. Often optimal decisions need to be taken in the presence of trade-offs between two or
more conflicting objectives, Stochastic Frontier Analysis; a method of economic modeling in which a stochastic component that describes random shocks affecting the production process is added. These shocks are not directly attributable to the producer or the underlying technology, etc. (in section Integration of computational models). There are also papers that document work done, especially in DEA, that have provided valuable sources of cross reference, which are discussed in the section, Referred Literature Survey.

1.3.1 Referred Literature Survey

A number of sources have provided a rich panorama of the literature and published work in our area of research.

Important among these are Allen et al. (1997), which provides a review of the evolution, development and future directions on the use of weight restrictions and value judgments in DEA. The paper concentrates on the implications of weight restrictions on the efficiency, targets and peer comparators of inefficient DMUs and indicates future research directions in this area. Angulo-Meza and Lins (2002) present a detailed review of methods for increasing discrimination between efficient DMUs in Data Envelopment Analysis; a non-parametric, LP technique used in the estimation of production functions and has been used extensively to estimate measures of technical efficiency in a range of industries, while Bauer (1990) discusses recent developments in the econometric approach to the estimation of stochastic frontiers such as production, costs, and profit functions, indicating areas requiring further work.

Emrouznejad et al. (2008) presents an extensive listing of DEA research covering theoretical developments as well as “real-world” applications from inception to the year 2007.

Some of the past accomplishments of DEA as well as some of its future prospects
are covered in Cooper et al. (2007b). The authors discuss and demonstrate relations between various DEA models and go on to indicate paths for future developments. Emrouznejad (2009) presents an extensive listing of DEA research covering theoretical developments as well as “real-world” applications from inception to the year 2007.

Thanassoulis et al. (2004) reviews various methodologies that have been used for incorporating value judgments in DEA while pointing out the main reasons why value judgments may be required in an efficiency analysis. Gonçalves et al. (2004) build on and formalize this approach, establishing the necessary conditions for an unobserved DMU to have an equivalent set of weight restrictions.

Jamasb and Pollitt (2001) focus on a survey of DEA applications to promote efficiency improvement in electricity transmission and distribution utilities. This paper discusses the main benchmarking methods and presents the findings of a survey of the methods used in the OECD and other countries. The authors outline the main outstanding issues and lessons for best practice implementation of benchmarking for regulation.

Liu et al. (2013) surveys the DEA literature and indicate papers playing the central role in DEA development. The authors also mention the latest active subareas within DEA.

Moving on to classification using ANN, Jain et al. (2000) summarize and compare some of the well-known methods used in various stages of a pattern recognition system and identify research topics and applications which are at the forefront of this field.
1.3.2 DEA

In their book, Charnes et al. (1996), the authors establish a reference source for DEA that presents an integrated framework for basic DEA models and their extension. It also provides insights into DEA usage and interpretations and highlights the range of potential uses of DEA through examples of novel applications.

In their paper, Park et al. (2012) propose a method of stepwise benchmarking based on three criteria: preference, direction and similarity. The first criterion, preference, is used for selecting an ultimate benchmark target; the second criterion, direction is used in selecting intermediate benchmark targets which are located more closely to the improving path; and the third criterion, similarity is used for determining intermediate benchmark targets which are similar to the DMU under evaluation. Considering these three criteria, the authors develop a method of constructing a more practical and feasible sequence of benchmark targets. This paper has provided valuable insights for the researchers’ paper Patel and Bose (2013a), “Seeking alternative DEA Benchmarks” which has been discussed in chapter 2.

Among the many papers based on DEA that have provided valuable insight to this research are Banker (1984), where the author develops the relation between the most productive scale size for specific input and output mixes and returns to scale for multiple-inputs multiple-outputs situations and uses this relation to extend the applications of DEA.

Banker et al. (1984) introduce a new variable which makes it possible to determine whether operations were conducted in regions of increasing, constant or decreasing returns to scale (in multiple input and multiple output situations). In their paper, Banker and Morey (1986b) evaluate technical and scale efficiencies of DMUs when some
of the inputs or outputs are exogenously fixed and beyond the discretionary control of DMU managers. This is further generalized and extended in Golany and Roll (1993). Cooper et al. (1996) formulate DEA models to include cases in which inputs and outputs are stochastic, as well as cases in which only the outputs are stochastic.

Podinovski (2001) discusses an alternative approach based on incorporating weight restrictions that avoids underestimating the maximum relative efficiency of DEA.

Cooper et al. (1999) extends DEA allowing imprecise data which are transformed to linear programming forms, with the widely used Assurance Region concepts. The resulting model, AR-IDEA includes imprecise data capabilities, assurance region and cone-ratio envelopment concepts.

Cooper et al. (2002) treat congestion in DEA, replacing stochastic models with their deterministic equivalents, and then identifying certain conditions that allows the resulting non-linear problems to be replaced by ordinary, deterministic DEA models.

There are many papers that provide innovative applications of DEA, among which are Fried et al. (1993); Hjalmarsson and Veiderpass (1992); Khalili et al. (2010); Kim (1987); Kumbhakar and Hjalmarsson (1997); Rebba and Rizzi (2003) and Shaneth et al. (2009). Taylor et al. (1997) use DEA/AR in their application of DEA to 48 large U.S. banks, while Zanakis et al. (2007) present a DEA output-oriented assurance-region model to assess the efficiency of 116 countries in battling the HIV/AIDS pandemic.

Stewart (2010) extend the standard data envelopment analysis (DEA) model to include longer term top management goals. The authors recognize the fact that benchmarking for DMUs should ideally include a component of future planning, and develop a goal programming structure to find points on the efficient frontier which are realistically achievable by DMUs.

Thanassoulis and Dyson (1992) develop models which can be used to estimate
Chapter 1. Introduction

alternative input-output target levels to render relatively inefficient organisational units efficient. models can incorporate preferences over potential improvements to individual input output levels so that the resultant target levels reflect the user’s preferences over alternative paths to efficiency.

Yu et al. (1996) study the structural properties of DEA efficient surfaces of the production possibility set under the Generalized Data Envelopment Analysis model. Wei and Yu (1997) study the properties of the K-cone in a generalized DEA model. The authors claim that the results of this paper will play an important role in determining the structure of efficient frontiers.

Zhu (2003) develop an extension to DEA that compares two different approaches in dealing with imprecise outputs and inputs, under a non-linear programming problem called “imprecise DEA” (IDEA). In their book, Zhu (2009), add several new DEA models and approaches. Some of the innovative DEA-based models discussed are a “Supply Chain Model”, models for “Two-stage Processes” and models with restricted multipliers.

1.3.3 Artificial Neural Network

Artificial Neural Networks (ANNs) have been widely used to not only solve a wide range of optimization problems but also to create new architectures. A topical introduction to this ability of ANN is provided in Cichocki and Unbehauen (1996); Skapura (1996); Livingstone (2008); Harvey (1994); Werbos (1988) and Zurada (1992). Hopfield and Tank (1985) discuss the general principles involved in constructing networks to solve specific problems. Results of computer simulations of a network designed to solve the well-defined Traveling-Salesman Problem are presented and used to illustrate the computational power of the networks. Lapedes and Farber (1987) develop a formalism
for non-linear signal processing using the backpropagation learning algorithm.

Cheng and Titterington (1994) outline areas of statistical interest, and provide examples of ANN models. The authors describe the ANN architectures and training rules and provide a statistical commentary.

Learning algorithms in neural networks has been the subject of much research. Atiya (1987) studies learning in a general network, while Rohwer (1990) proposes a method for training the dynamical behavior of a neural network that is applicable to discrete-time networks with arbitrary feedback. Learning in recurrent neural networks is studied in Fang and Sejnowski (1990); Gherrity (1989); Ackley et al. (1985); Pearlmutter (1988) and Gori et al. (1989). Lockery et al. (1990) demonstrates a model for dynamic networks using recurrent backpropagation. Ackley et al. (1985) describes a parallel search algorithm based on Boltzmann machines. Boltzmann-based models with asymmetric connections are investigated in Allen and Alspector (1990), where the authors show that though they are initially unstable, these networks spontaneously self-stabilize as a result of learning.

In the book Haykin (1999) discusses in depth the learning process, back propogation, radial basis functions, recurrent networks, self-organizing systems, and other aspects of ANN. The author covers a wide range of ANN applications, including support vector machines and reinforcement learning/neurodynamic programming. A number of case studies illustrate real-life, practical applications of neural networks.

Chen et al. (1992) presents a method for improving the performance of ANNs for linear and nonlinear programming by proposing a new combination penalty function which can ensure that the equilibrium point is acceptably close to the optimal.

An application of ANN for pricing of options using three different ANN models is described in Samur and Tekin Temur (2009).
1.3.4 Integration of computational models

A very exciting and potentially rich area is that of integration of technologies. Azadeh et al. (2006, 2011) integrate ANN and fuzzy-DEA to come up with the non-parametric efficiency frontier and identify optimum location of solar-plants with possible noise, non-linearity, complexity and environmental uncertainty.

An approach derived from the integration of DEA data envelopment analysis and a multi-attribute value function is presented in Belton and Vickers (1993). The authors claim that this is particularly well-suited to analyze the efficiency of a small number of units.

Costa and Markellos, 1997 present a method integrating ANN along with DEA and corrected least squares (COLS), to set out comparative annual efficiency measures for the London Underground, for the period 1970 to 1994.

Deprins et al. (1984) define and compare three different methods of measuring technical efficiency in an application to Belgian postal data. The first one is that of adjusting a Cobb-Douglas production frontier; the second is that of computing the convex hull of the data, à la Farrell; the third one is developed on the basis of the sole assumptions of input and output disposability.

Golany (1988) presents a multi-objective linear-programming procedure to aid decision-makers in setting up goals for desired outputs. The procedure described relies on empirical production functions generated by the use of DEA and presents the decision-maker with a set of alternative efficient points.

In their paper, Cheng and Titterington (1994) points out some of links of ANN with statistical methodology and encourages cross-disciplinary research in the directions most likely to bear fruit. The authors cite areas of statistical interest and a series of
examples indicating the flavor of ANN models.

A two-phase model is discussed in Cook and Bala (2007), where in the first phase a classification tool is used to quantify a functional relation for the performance model leading to an orientation of the variables in terms of “inputs” or “outputs” that defines the second-phase DEA model.

There is a wealth of literature integrating cluster analysis and Data Envelopment Analysis; a non-parametric, LP technique used in the estimation of production functions and has been used extensively to estimate measures of technical efficiency in a range of industries. Some of the works that have inspired and motivated this research are as described below.

Po et al. (2009) proposed a DEA-based clustering approach that employs the piecewise production functions derived from the DEA method to cluster the data with input and output items. Thus, each evaluated decision-making unit (DMU) knows the cluster that it belongs to and checks the production function type that it confronts.

Hirschberg and Lye (2001) applies cluster analysis to the results of a Data Envelopment Analysis of productive behaviour to identify groups among a set of different objects.

There are several other instances of integration of DEA and cluster analysis. Among them are Krishnapuram and Keller (1993); Tsao et al. (1994); Wu and Yang (2002). Tsao et al. (1994) uses a fuzzy Kohonen clustering network integrating the fuzzy c-means model into the learning rate and updating strategies of the Kohonen network. In this paper, Yang (1993) survey fuzzy clustering in three categories – the first category is the fuzzy clustering based on fuzzy relation; the second is the fuzzy clustering based on objective function; and finally, the author gives an overview of a nonparametric classifier – the fuzzy generalized k-nearest neighbor rule.
Chapter 1. Introduction

Yet another example of integration is provided by Jolion et al. (1991). The author discusses a clustering algorithm based on a recently introduced statistical technique – the minimum volume ellipsoid. The MVE estimator identifies the least volume region containing $h$ percent of the data points – with the value of $h$ iteratively decreasing from 0.5.

Joro et al. (1998) leverage MOLP to provide interesting extensions to DEA and insist on using DEA and MOLP not as substitutes of each other, but rather as complements. Multi Criteria Decision Making (MCDM) is used along with DEA in Agrell and Tind (1998).

Stewart (1996) contrasts and compares DEA and MCDM highlighting the assumptions needed during model formulation in order to constrain weight flexibility in DEA.

1.4 Research Objective

The thesis draws from the review of literature, discussed in the last section. While the existing literature survey and literature about integration of DEA / ANN methods has provided motivation and the inspiration for this research, the reviewed literature about DEA has provided knowledge about subtleties of DEA technique that are leveraged in this research.

The objective of this research is to refine the DEA benchmarking process, addressing issues that are related to data-sensitivity. Sample size as well as the input-output characteristics of the DMUs significantly influence the efficiency computations, as also the peer benchmarks identified for inefficient DMUs. The aim of our research is to extend and refine DEA so as to produce consistent and meaningful results even when the sample characteristics are unfavourable.
Chapter 1. Introduction

- To reduce sensitivity associated with conventional DEA measures of efficiency
  DEA efficiency measures and benchmarks depend to a great extent on the DMUs involved. The strengths of individual DMUs in terms of utilization of specific inputs or production of specific outputs influence the overall analysis. This is more so when the DMUs that are “leaders” are few in number and share input-output characteristics that are different from the other, less efficient DMUs. In such conditions, conventional DEA measures and benchmarks often misrepresent the underlying reality and provide benchmarks that are not very representative.
  
  This research attempts to reduce the dependence of DEA on the characteristics of the sample DMUs, and in particular the influence of input-output characteristics of the “leader” DMUs.

- To approach benchmarking from feasibility angle, so that DMUs can reasonably achieve the benchmark goals

  An important management outcome of DEA is the ‘benchmarks’ evaluated for inefficient DMUs. DEA provides measures of efficiency which quantify the degree of inefficiency a DMU has; additionally it prescribes benchmarks which may be emulated by an inefficient DMU in order to attain efficiency. However, this theoretical exercise often sets a tall order for an inefficient DMU - the emulation of the benchmark is often infeasible.

  In this research we try to provide a “performance improvement path” for an inefficient DMU that provides meaningful, feasible targets in small time steps.
1.5 Research Methodology

In this research we have created three separate inter-disciplinary frameworks that can ameliorate the problems of infeasible benchmarks and unrepresentative efficiency measures. The frameworks extend DEA by looking at different distance measures to the efficient frontier and incorporating clustering to identify “similar” DMUs that can be meaningfully compared. We also utilize Artificial Neural Networks to identify an extended, and arguably more representative efficiency frontier.

The research primarily focuses on framework building and we argue analytically why each framework is suitable. We also use different secondary data sets to empirically illustrate the efficacy of our methods. These data are from
1. Twelve hospitals data set from Cooper et al. (2007a)
2. Data from Market (2011) for 99 Indian Microfinance Institutions that have reported results for 2011
3. Data from Cooper et al. (2007a) for public libraries in 23 wards of the Tokyo Metropolitan Area in 1986