Chapter 2

LITERATURE REVIEW

This chapter reviews literature in the areas of productivity and performance measurement using mathematical programming and multi-level planning for resource allocation and efficiency measurement. The review is described in the following sections.

i) Data Envelopment Analysis

ii) Multi-Level Programming

iii) Goal Programming

iv) Goal Programming and Data Envelopment Analysis

v) Decision Making In A Fuzzy Environment

The chapter is organized as follows. First, definitions and concepts of efficiency are presented. This is followed with a presentation of the fundamental concepts and constructs of DEA. Then work in performance and efficiency measurement with DEA is summarized. Second, the multi-level programming approach is reviewed. The specific case of the bilevel programming model is outlined. Third, the goal programming formulation is presented. Fourth, a formulation integrating goal programming and DEA (GoDEA) is presented. Finally, decision-making in a fuzzy environment is reviewed with a brief note on fuzzy set theory and the use of fuzzy set theory in goal programming and data envelopment analysis.

2.1 DEFINITIONS AND CONCEPTS

In the management science literature, productivity and performance measurement have traditionally been concerned with some factors (inputs and outputs), processes, or machines rather than the organizational whole. For example, one measurement technique is to calculate the ratio of total output to a particular input *i.e.*, partial factor productivity. The most common measure is that of labor productivity (*e.g.*, output per man-hours) while another common measure is capital productivity (*e.g.*, rate of return on capital utilized) (Stainer (1997)). According to Stainer (1997), such ratios face a fundamental problem wherein external factors may affect their computation and have no relationship to efficient resource usage.

Productivity research led to the development of other measures that incorporated all the important factors in aggregated form. These measures offered more insight about technical and financial performance of an organization. The concept of technical efficiency introduced by Farrell (1957) is a result of these concerns. Charnes *et al*. (1978) further extended Farrell's (1957) work and developed a mathematical programming approach to measure relative efficiency of decision making units. Basic concepts and their definitions are summarized below.

2.1.1 Production Technology

A *production technology* is defined as the set (X, Y) such that inputs $\boldsymbol{X} = (x_1,x_2,\ldots x_i) \in \mathbf{R}^i_+$ are transformed into outputs $\boldsymbol{Y} = (y_1,y_2,\ldots y_j) \in \mathbf{R}^j_+$. Färe *et al.* (1994) describe production technology with the following notation.

L(y) is the input set such that:

$$L(y) = \{x: (y,x) \text{ is feasible}\}, \tag{2.1}$$

$\forall$ y $\in$ R^s_+ $\exists$ an isoquant *Isoq*L(y) such that:

$$IsoqL(y) = \{x: x \in L(y), \lambda x \notin L(y), \lambda \in [0,1)\}, \tag{2.2}$$

and an efficient subset *Eff*L(y) such that:

$$EffL(y) = \{x: x \in L(y), x' \notin L(y), x' \leq x\}^1 \tag{2.3}$$

2.1.2 Production Function

A *production function* is defined as the relationship between the outputs and inputs of a production technology. Mathematically, a production function relates the amount of output (Y) as a function of the amount of input (X) used to generate that output. Technical efficiency (Section 2.2) is assumed for a production function *i.e.*, every feasible combination of inputs generates the maximum possible output or all outputs are produced using the minimum feasible combination of inputs. For example, the production function for input X and output Y is:

$$Y = f(X) \tag{2.4}$$

2.1.3 Isoquant

An *isoquant* is defined as the locus of points that represent all possible input-output combinations that defines the production function for a constant level of output (or a constant level of input). Each point on the isoquant represents a unique production technology. For example, the isoquant (input orientation) for output level Y^0 (a specific realization of output Y in section 2.1.2 above) from input X is:

---

[1] $x_i' \leq x_i$, $\forall$ $i = 1,\ldots,m$ and $x_i' < x_i$,for at least one component i.

$$Y^0 = g(X) \tag{2.5}$$

This isoquant is shown in Figure 2.1. In this case, further the isoquant is from the origin in the positive quadrant, the greater is the output level.

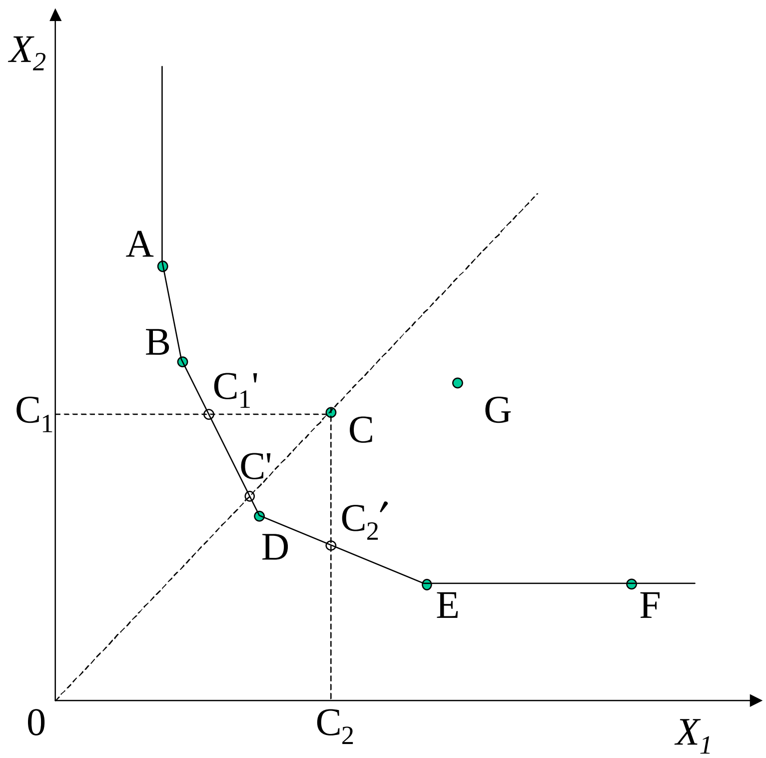

Figure 2.1 Isoquant for Output Level Y^o

2.1.4 Radial and Non-Radial Measures of Efficiency

Consider the isoquant in Figure 2.1 for output level Y^0. Suppose there are firms A, B, C, D, E, F, and G, each of which produce the same output Y^0 consuming inputs X_1 and X_2. Firms A, B, D, E, and F are the firms consuming the least amount of each input to produce Y^0. Therefore, these firms define the isoquant and also lie on it. Firms C and G consume more inputs to generate Y^0, are enveloped by the isoquant and are therefore inefficient. The isoquant thus serves as the standard of comparison for the firms. This is the essence of the concept of relative efficiency and is explained in detail in Section 2.2. There are two ways to measure the efficiency for a given firm, (i) radially and (ii) non-radially.

Consider the inefficient firm C. Let C' be a virtual firm that is the convex combination of firms B and D. C and C' lie on the same ray through the origin and C' lies on the isoquant. Therefore, the radial measure of technical efficiency for C is:

$$\mathrm{TE}_{\mathrm{Radial}}\ (\mathrm{C}) = \frac{\mathrm{OC}'}{\mathrm{OC}} \tag{2.6}$$

This means that for C to become efficient it must operate at C''s input levels *i.e.*, C must *radially* or equi-proportionately reduce its inputs to C''s levels. However, an equi-proportional reduction in inputs may not always be feasible. In this case, the non-radial measure of efficiency is more appropriate. The non-radial measure of technical efficiency for C for inputs X_1 and X_2 is:

$$\mathrm{TE}_{\mathrm{Non\text{ - }Radial}}\ (\mathrm{C}) = \frac{\mathrm{C_1C'_1}}{\mathrm{C_1C}} \tag{2.7}$$

$$\mathrm{TE}_{\mathrm{Non\text{ - }Radial}}\ (\mathrm{C}) = \frac{\mathrm{C_2C'_2}}{\mathrm{C_2C}} \tag{2.8}$$

Here, the inputs are reduced individually by different proportions (non-radially) to reach the efficient subset ABDE while maintaining the same level of output and not altering the input levels of the remaining inputs. Therefore, separate efficiency scores are obtained for each input.

2.1.5 Returns to Scale

In production theory the change in output levels due to changes in input levels is termed as returns to scale. Returns to scale can be *constant* or *variable*. Constant returns to scale (CRS) implies that an increase in input levels by a certain proportion results in an increase in output levels by the same proportion. Figure 2.2 shows this linear relationship between the inputs and outputs. Variable returns to scale (VRS) implies that an increase in the input levels need not necessarily result in a proportional increase in output levels *i.e.,* the output levels can increase (increasing returns to scale) or the output levels can decrease (decreasing returns to scale) by a different proportion than the input increment.

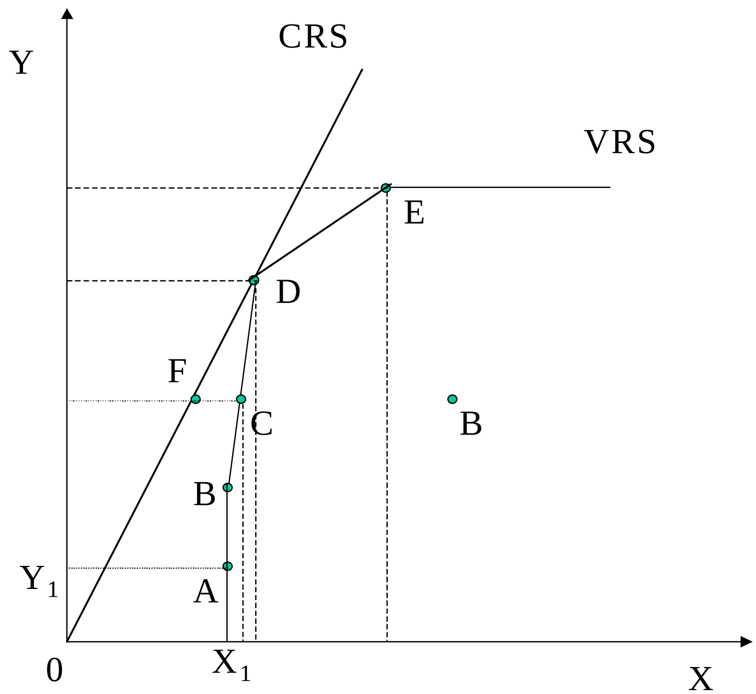

Figure 2.2 Constant and Variable Returns to Scale

Geometrically, this means that the linear relationship between inputs and outputs in the case of CRS is replaced by a curve with a changing slope. Figure 2.2 shows the piece wise linear curve with varying slopes. As the slope of the curve increases the production technology displays increasing returns to scale (*e.g.,* from B to D). Where the slope of the curve decreases the production technology displays decreasing returns to scale. (*e.g.,* from D to E). And where the curve has a zero slope (from point E to ∞) the production technology experiences no increase in output for any further increase in input. Where the curve has a zero slope (from X_1 to A) the output jumps from 0 to Y_1 for an input usage of X_1.

2.1.6 Definitions of Technical Efficiency

The concepts presented above enable the discussion of the two definitions of technical efficiency that are reported in the literature (Färe and Lovell (1978)). The first one is the radial definition presented by Debreu (1951) and Farrell (1957). The input-reducing radial measure of technical efficiency for a unit is defined as the difference between unity (100% efficiency) and the maximum equi-proportional reduction in inputs (while maintaining the production of originally specified output levels). If this difference is zero then the unit is efficient else it is inefficient. The output-increasing radial measure of technical efficiency is defined as the difference between unity (100% efficiency) and the maximum augmentation of outputs (while still utilizing the originally specified input levels). Again, the unit is efficient if this difference is zero else it is inefficient.

The second definition is Koopmans' (1951) definition of technical efficiency. The firm is technically efficient if and only if an increase in one output results in a decrease in another output so as not to compromise the input level or else results in the increase of at least one input. Stated otherwise, the definition implies that a decrease in one input must result in an increase in another input so as not to compromise the output, or else must result in the decrease of at least one output.

The difference between the two definitions is explained through Figure 2.1. The radial definition provided by Debreu (1951) and Farrell (1957) terms all firms on the isoquant with output level Y^0 as efficient. However, Koopman's (1951) non radial-definition deems firm F as inefficient. This is due to the fact that though firm F lies on the isoquant it does not lie on the efficient subset of the isoquant defined by ABDE. In other words, E produces the same output with fewer inputs (lesser amount of x_1) than F, and therefore F is inefficient. This case highlights the situation where a unit may lie on the isoquant but still consume excess inputs compared to other units on the isoquant. The next section provides a detailed discussion of technical efficiency.

2.2 TECHNICAL EFFICIENCY

Traditionally, labor productivity was considered as an overall measure of efficiency (Farrell (1957)). According to Farrell (1957) this ratio (*e.g.*, units produced divided by labor hours) was inappropriate as a measure of technical efficiency (TE) as it incorporated only labor and ignored other important factors such as materials, energy, and capital. Thus Farrell (1957) proposed a measure of *technical efficiency* that incorporated all inputs in an aggregated scalar form and also overcame the difficulty of converting multi-component input vectors into scalars. Thus the technical efficiency formulation for multiple input-output configurations is:

$$\text{TE} = \frac{\text{Aggregate Output Measure}}{\text{Aggregate Input Measure}} \tag{2.9}$$

The inputs are all resources that are consumed to generate the outputs. From equation (2.9) it can be seen that technical efficiency for a firm relates to its ability to:

(i) produce maximum outputs for a constant input usage (output-increasing efficiency, or

(ii) use minimum inputs to generate a constant output production (input-reducing efficiency).

Technical efficiency measurement generally involves comparing a decision making unit's (DMU's) production plan to a production plan that lies on the efficient production frontier or isoquant (Fried *et al.* (1993), Färe *et al.* (1994), Charnes *et al.* (1994)). As presented in section 2.1.1, a production plan for a DMU represents its input usage and output production. The concept of a production plan motivates two types of technical efficiency measurement, input-reducing and output-increasing. Input-reducing efficiency refers to the production of a constant output set while reducing the level of inputs used to the least possible. Output-increasing efficiency refers to maintaining a fixed level of inputs while producing the maximum possible set of outputs.

The notion of comparisons of production plans leads to the need for deriving a "standard of excellence" to serve as a benchmark. This standard must represent that level of technical efficiency that is achieved with (i) the least amount of inputs and constant outputs (for input-reducing efficiency) and (ii) the maximum production of outputs with constant inputs (for output-increasing efficiency). The literature reports three approaches to measure technical efficiency: (i) the index numbers approach (ii) the econometric approach, and (ii) the mathematical programming approach. The index numbers approach includes multi-factor productivity models, financial and operational ratios (Parkan (1997)). The econometric approach presupposes a theoretical production function to serve as the standard of technical efficiency. The Cobb-Douglas, Translog, and Leontief type functions are most commonly used to approximate the production function as they are easily transformed into linear forms. Econometric models are further divided into *deterministic* and *stochastic* models. For a detailed discussion of the econometric approach the reader is referred to Girod (1996) and Lovell (1993). The mathematical programming approach does not require the use of a specified functional form for the production data. This approach was pioneered by Charnes, Cooper and Rhodes (1978) and is called Data Envelopment Analysis (DEA). DEA is defined by Giokas (1997) as follows:

> *"DEA measures relative efficiency [of DMUs] by estimating an empirical production function which represents the highest values of outputs/benefits that could be generated by inputs/resources as given by a range of observed input/output measures during a common time period."*

While econometric methods (*e.g.,* regression analysis) employ "average observations" mathematical programming methods (*e.g.,* DEA) use "production frontiers" or "best practice observations" for efficiency analysis. A detailed discussion of input-reducing and output-increasing orientations of technical efficiency and DEA is provided in the subsequent sections.

2.2.1 Input-Reducing and Output-Increasing Orientations of Technical Efficiency

The concepts of input-reducing and output-increasing orientations of Farrell's (1957) technical efficiency measure are presented through the following example. Consider a decision-making unit (DMU) that uses $i = 1, 2, \ldots, \boldsymbol{I}$ inputs to produce $j = 1, 2, \ldots, \boldsymbol{J}$ outputs. Let the matter of interest be DMU productivity performance over time (say, one year or twelve months) *i.e.*, let each month represent a DMU. Denote the input vector for the n^{th} month as $X_n = [x_{in}]$ and the output vector for the n^{th} month as $Y_n = [y_{jn}]$ (Figure 2.3).

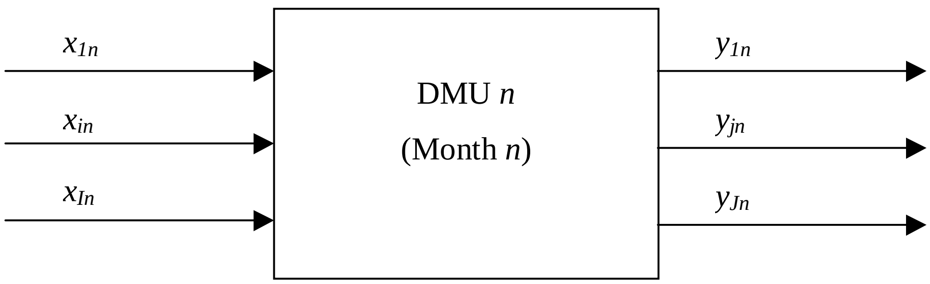

Figure 2.3 Input and Output Vectors for the n^{th} Month

It is assumed that each component of the input and output vectors is uniquely identifiable and quantifiable (Hoopes and Triantis (1999)). Thus, the objective for the DMU (month) would be to minimize usage of each input resource and maximize the production of each output type. Farrell (1957) defined the "*efficient*" transformation of inputs into outputs as the *efficient production frontier* or *isoquant*. An isoquant can be oriented for input-reduction or output-augmentation.

An input-reducing isoquant is defined by the observations that are efficient relative to the other observations in the data set. This isoquant represents the minimum input usage that is required to produce a constant set of outputs. Points on the isoquant are observations with different input mixes producing the same level of outputs. Such an isoquant is assumed to be convex to the origin and to have a negative slope. The convexity assumption allows for *virtual* production plans that are obtained as a weighted combination of *actual* production plans.

Strong disposability of inputs i.e., an increase in any input keeping other inputs constant must result in an increase in the level of outputs. In other words, given the negative slope assumption, different input mixes can be obtained without compromising the output level. With strong disposability of inputs an increase in any input without change in other inputs must result in the observation moving to a higher isoquant. If the output level remains constant than it would imply weak disposability of inputs.

To illustrate an input-reducing isoquant consider a production plan with a constant output set Y^0 and two inputs $X = [X_1, X_2]$. Let the production technology exhibit constant returns to scale where an increase in inputs results in an equi-proportional increase in the output. The resultant input-reducing isoquant then represents the output level. For every output level there exists an input-reducing isoquant. Therefore, the input-reducing isoquant is a function of the output level. In Figure 2.4 SS' represents the input-reducing isoquant for output level Y^0 attained with inputs $X = [X_1, X_2]$. The case of multiple inputs (greater than two) is extended similarly.

An output-reducing isoquant or production possibility frontier is constructed similarly by observations deemed relatively efficient in the data set. This frontier represents the maximum output production possible with consumption of constant inputs. Efficient observations are points on the frontier with different output mixes produced with the same level of inputs. Once again, let the production technology exhibit concavity with respect to the origin, negative slope, and constant returns to scale. The concavity property permits observations obtained as weighted combinations of actual observations. The negative slope permits *strong disposability of outputs i.e.,* a decrease in any one output keeping all other outputs constant must result in the decrease in the level of inputs. Therefore different output mixes can be obtained without compromising the level of inputs. The constant returns to scale assumption implies that a decrease in outputs results in an equi-proportional decrease in inputs. The strong disposability of outputs ensures that a decrease in any output without change in other outputs must result in the observation moving to a lower frontier. If the input level remains the same then it implies weak disposability of outputs.

To illustrate an output-increasing frontier consider a production plan with two outputs $Y = [Y_1, Y_2]$ and a constant input set X^0. The resultant output-increasing frontier then represents the input level. For every input level there exists an output-increasing frontier. Therefore, the output-increasing frontier is a function of the input level. In Figure 2.5 RR' represents the output-increasing frontier for input level X^0 attained with outputs $Y = [Y_1, Y_2]$. The case of multiple outputs (greater than two) is extended similarly.

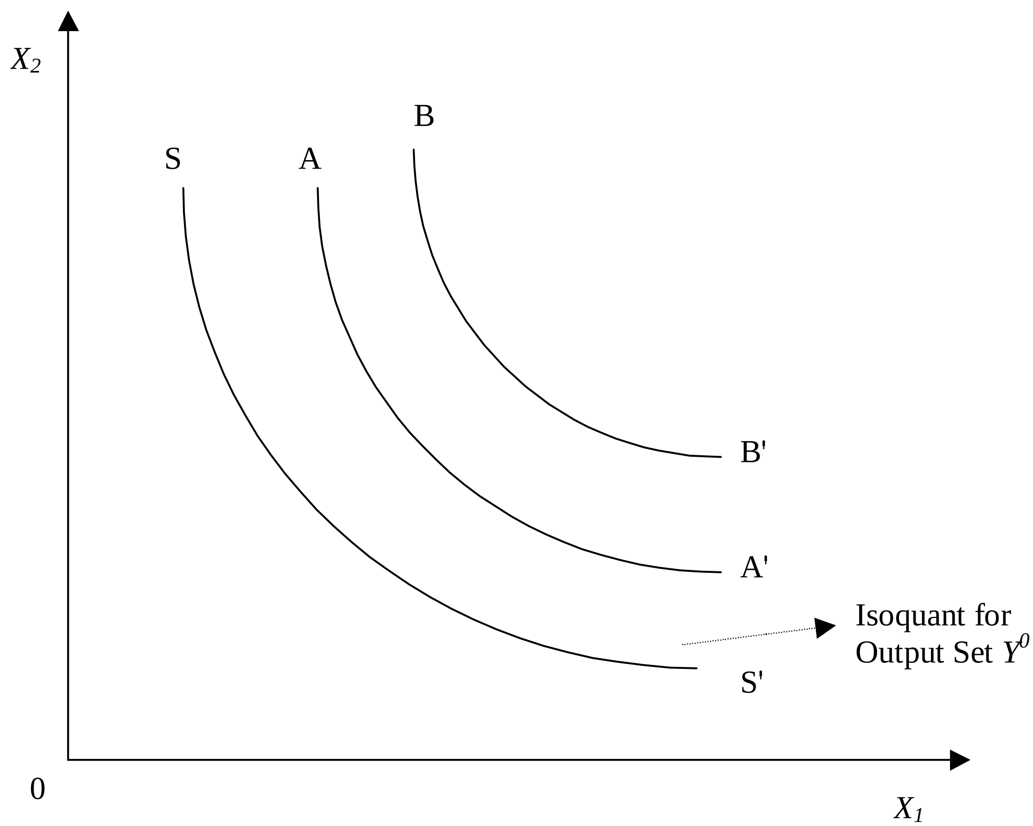

Figure 2.4 Input-Reducing Isoquant Orientation

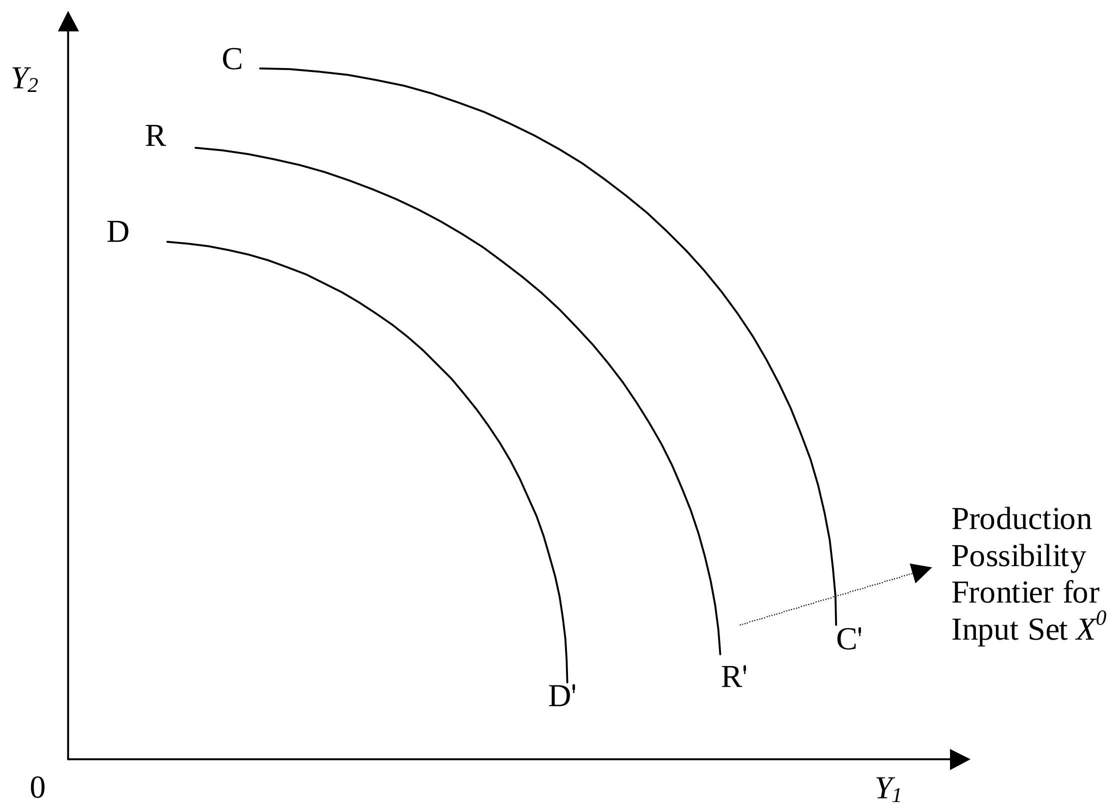

Figure 2.5 Output-Increasing Isoquant Orientation

It should be noted that while the isoquant constructs are obtained from input/output levels, the production function is unique for a data set. In other words, isoquants are particular realizations of a production function for a given input/output orientation. Farrell (1957) uses the isoquant construct(s) presented above as the efficient frontier to compare performance of different observations. Since the isoquant is constructed from observed data, the relative comparison of observations is based on the Pareto-Koopmans condition or Pareto optimality condition. For the input-reducing case this condition is stated by Charnes *et al.* (1978, p.13) as:

> *"If, for a given observation's input-output mix, it is not possible to find an observation or combination of observations that produce the same amount of output with less of some input and no more of other inputs, then the given observation is efficient. Otherwise, the given observation is inefficient."*

The Pareto optimality concept is illustrated graphically as follows. Figure 2.6 presents the input-reducing case. For graphical simplicity, suppose that all DMUs utilize two inputs X_1 and X_2 to produce a constant output set Y^0 *i.e.*, the output space is Y^0. Here, DMUs A, C, D, and E are efficient as they lie on the efficient frontier (or isoquant for output level Y^0) while DMU B is inefficient as it lies *above* the isoquant. Let B' be a convex combination of DMUs C and D. Then, both DMU B and virtual DMU B' produce the same output set Y^0, but DMU B consumes more resources than virtual DMU B'. Therefore, the input-(in)efficiency score of DMU B is given as:

$$TE_{Input}(B) = \frac{OB'}{OB} \tag{2.10}$$

Similarly, Figure 2.7 presents the output-increasing case. Again, for graphical simplicity, suppose that all DMUs produce two outputs Y_1 and Y_2 from a constant input set X^0 *i.e.*, the input space is X^0. Here, DMUs F, G, H, and I are efficient as they lie on the frontier for input level X^0 while DMU J is inefficient as it lies *under* the frontier. Let J' be a convex combination of DMUs G and H. Then, both DMU J and virtual DMU J' consume the same amount of inputs, but virtual DMU J' produces more output than DMU J. Therefore, the output-(in)efficiency score of DMU J is given as:

$$\mathrm{TE}_{\mathrm{Output}}\,(\mathrm{J}) = \frac{\mathrm{OJ}}{\mathrm{OJ'}} \tag{2.11}$$

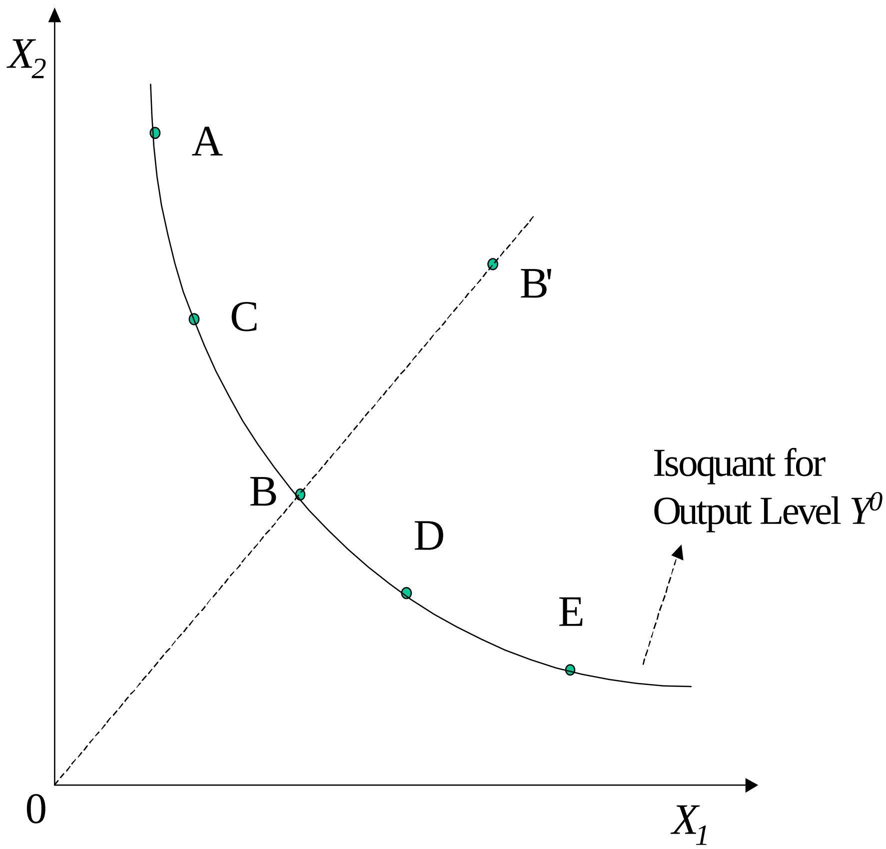

Figure 2.6 Input-Reducing Technical Efficiency

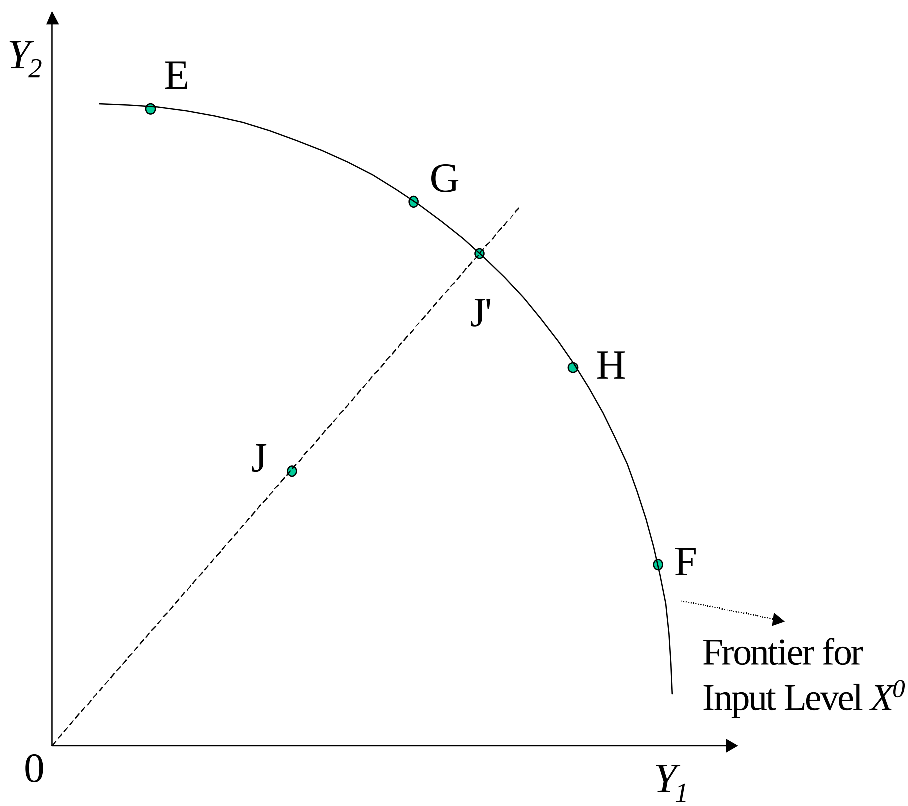

Figure 2.7 Output-Increasing Technical Efficiency

The above review of Farrell's (1957) technical efficiency measure clearly outlines the importance of the assumptions of convexity, negative slope and constant returns to scale which define the production function. Econometric methods theoretically fit a function to the production technology. Aigner and Chu (1968) first applied the Cobb-Douglas function to estimate the efficient production frontier. However, it has been observed that theoretically derived functions provide inaccurate approximations of the production technology as the complexity of the technology increases. This prompted researchers to look closely at the mathematical programming methods that empirically derive the efficient production frontier from observed data. Charnes, Cooper, and Rhodes (1978) extended Farrell's (1957) work in the measurement of technical efficiency and developed Data Envelopment Analysis (DEA). The DEA methodology allows the relaxation and the enhancement of some of Farrell's (1957) assumptions for the production function and the production technology. The DEA review is presented at this time.

2.3 DATA ENVELOPMENT ANALYSIS (DEA)

Data Envelopment Analysis (DEA) is a non-parametric performance assessment methodology originally designed by Charnes, Cooper and Rhodes (1978) to measure the relative efficiencies of organizational or decision making units (DMUs). The DEA approach applies linear programming techniques to observed inputs consumed and outputs produced by decision-making units and constructs an efficient production frontier based on best practices. Each DMU's efficiency is then measured relative to this frontier. In other words, DEA assesses the efficiency of each DMU relative to all the DMUs in the sample, including itself. This relative efficiency is calculated by obtaining the ratio of the weighted sum of all outputs and the weighted sum of all inputs. The weights are selected so as to achieve Pareto optimality for each DMU. The DEA methodology is concerned with *technical efficiency i.e.,* the physical levels of outputs produced and inputs consumed as compared to *allocative efficiency i.e.,* the optimal input mix given input prices, and *price efficiency i.e.,* the optimal output mix given output prices (Lewin and Morey (1981)).

An appealing aspect of DEA is that it allows analysis of multiple-input multiple-output production technologies without requiring price or cost data. Also, the various input and output factors need not have the same measurement units *i.e.,* DEA is invariant to scaling of variables. This is important in public sector organizations where financial and cost data is often unavailable for all factors.

The DEA methodology helps to identify inefficient DMUs as well as the sources and amounts of inefficiency of inputs and/or outputs. The DEA formulation can incorporate both input-reducing and output-augmenting orientations as well as constant and variable returns to scale. The following discussion presents only the input-reducing orientation. The output-increasing orientation is analogous and derived similarly. However, different results are obtained from the two orientations under the variable returns to scale assumption (Färe and Lovell (1978)).

Since its original development, DEA has expanded considerably. Seiford (1996) has reported more than 800 references on the subject. Various applications of DEA to public organizations such as schools, banks, hospitals, armed services, shops, and local authority departments have been published. In this review, the foundations of the DEA framework, and the important formulations (input-reducing orientation) are presented.

2.3.1 The CCR Model

The DEA model originally proposed by Charnes, Cooper, and Rhodes (1978) is called the CCR model. This model allows input-reducing and output-increasing orientations and assumes constant returns to scale.

The CCR model is an extension of Farrell's (1957) classical work on technical efficiency. The DEA model requires complete information on inputs and outputs for a set of homogenous DMUs. The model is a fractional linear program that compares the efficiency of each DMU with all possible linear combinations of the other DMUs (including the one under consideration).

In mathematical terms, consider a set of n DMUs, where DMU j has a production plan (Xj, Yj) with $\boldsymbol{X_j} = (x_1, x_2, \ldots, x_m)$ inputs and $\boldsymbol{Y_j} = (y_1, y_2, \ldots, y_s)$ outputs. Let $\boldsymbol{U} = (u_1, u_2, \ldots, u_m)$ and $\boldsymbol{V} = (v_1, v_2, \ldots, v_s)$ be weight vectors. Let the variables be defined as:

c = DMU whose technical efficiency is being measured
x_{ik} = quantity of input i consumed by DMU k
y_{jk} = quantity of output j produced by DMU k
u_i = weight assigned to input i
v_j = weight assigned to output j
ε = very small positive number

The CCR model is then written as:

Ratio Form of the CCR Model (M1)

$$Max \quad \frac{\sum_{j=1}^{s} v_j y_{jc}}{\sum_{i=1}^{m} u_i x_{ic}} \tag{2.12}$$

$$subject\ to \quad \frac{\sum_{j=1}^{s} v_j y_{jk}}{\sum_{i=1}^{m} u_i x_{ik}} \le 1, \qquad k = \{1,2,\ldots,n\} \tag{2.13}$$

$$u_i \ge \varepsilon, \qquad i = \{1,2,\ldots,m\} \tag{2.14}$$

$$v_j \ge \varepsilon, \qquad j = \{1,2,\ldots,s\} \tag{2.15}$$

From Model M1 the efficiency of DMU *c* is measured as a weighted sum of outputs divided by a weighted sum of inputs. This efficiency is maximized subject to the efficiencies of all units and bounded above by 1. However, the key feature of the DEA model is that the weights ***U*** and ***V*** are not fixed exogenously, but are chosen (by the model) so as to maximize the efficiency of the DMU under consideration in comparison to the other DMUs which must also carry the same weights. In other words, the weights are so chosen that each DMU is shown in the best possible light. It is important to note that the weights will not necessarily be the same for each DMU. This "biased" choice of weights is summarized by Boussofiane, Dyson, and Thannasoulis (1991) (pp. 2) as both a strength and a weakness:

> *"It is a weakness because a judicious choice of weights may allow a unit [DMU] to be efficient but there may be concern that this has more to do with the choice of weights than any inherent efficiency. This flexibility is also a strength, however, for if a unit [DMU] turns out to be inefficient even when the most favorable weights have been incorporated in its*

efficiency measure then this is a strong statement and in particular the argument that the weights are inappropriate is not tenable."

The fractional linear program (**M1**) can be written as a linear program with $s + m$ variables and $n + s + m + 1$ constraints. The problem is then formulated as:

CCR Linear (Primal) Model (M2)

$$Max \quad \sum_{j=1}^{s} v_j y_{jc} \tag{2.16}$$

$$subject\ to \quad \sum_{i=1}^{m} u_i x_{ic} = 1 \tag{2.17}$$

$$\sum_{j=1}^{s} v_j y_{jk} - \sum_{i=1}^{m} u_i x_{ik} \leq 0, \qquad k = \{1,2,\ldots,n\} \tag{2.18}$$

$$-u_i \leq -\varepsilon, \qquad i = \{1,2,\ldots,m\} \tag{2.19}$$

$$-v_j \leq -\varepsilon, \qquad j = \{1,2,\ldots,s\} \tag{2.20}$$

Thus, it follows that **(M2)** must be solved for each DMU in turn. For computational ease DEA models are generally solved using the dual representation instead of the primal. The primal form has as many constraints as there are DMUs. The dual has as many constraints as there are inputs and outputs. In most cases the number of DMUs is much greater than the number of inputs and outputs. Therefore, the dual problem is smaller in size and more easily solved.

The dual of the CCR model **(M2)** is:

CCR Dual Model (M3)

$$Min \qquad \theta_c \tag{2.21}$$

$$subject\ \ to \qquad \theta_c x_{ic} - \sum_{k=1}^{n} z_k x_{ik} \geq 0, \qquad i = \{1,2,\ldots,m\} \tag{2.22}$$

$$\sum_{k=1}^{n} z_k y_{jk} \geq y_{jk}, \qquad j = \{1,2,\ldots,s\} \tag{2.23}$$

$$\theta_c\ , z_k \geq 0, \qquad k = \{1,2,\ldots,n\} \tag{2.24}$$

where,

θ_c = radial measure of technical efficiency

z_k = activity levels associated with inputs and outputs of DMU k

The optimal solution to the above problem, denoted as θ_c^*, is the degree of *input*-efficiency of DMU_c. A new weight vector $\boldsymbol{z} = (z_1, z_2, \ldots, z_k)$ appears in the dual formulation. This weight vector is unique for each DMU. The z_k's are the activity levels and characterize the level of performance of an efficient *virtual* DMU_c' against which the performance of DMU_c is compared. The dual seeks to find values of z_k so as to construct a composite (virtual) unit DMU_c' with outputs $\boldsymbol{\Sigma}\ z_k y_k$ and inputs $\boldsymbol{\Sigma}\ z_k x_k$ that outperforms DMU_c.

If both DMU_c and DMU_c' are found to perform at the same level then DMU_c is considered to be efficient and designated an input-efficiency score of one. In other words DMU_c will be efficient when it proves impossible to construct a virtual unit that outperforms it. If DMU_c utilizes more inputs than DMU_c', then DMU_c is considered inefficient and given an input-efficiency score less than one. This is so because it is possible for DMU_c' to produce the same output using lesser input than DMU_c. In this case, the optimal values of z_k will construct a virtual unit that outperforms DMU_c.

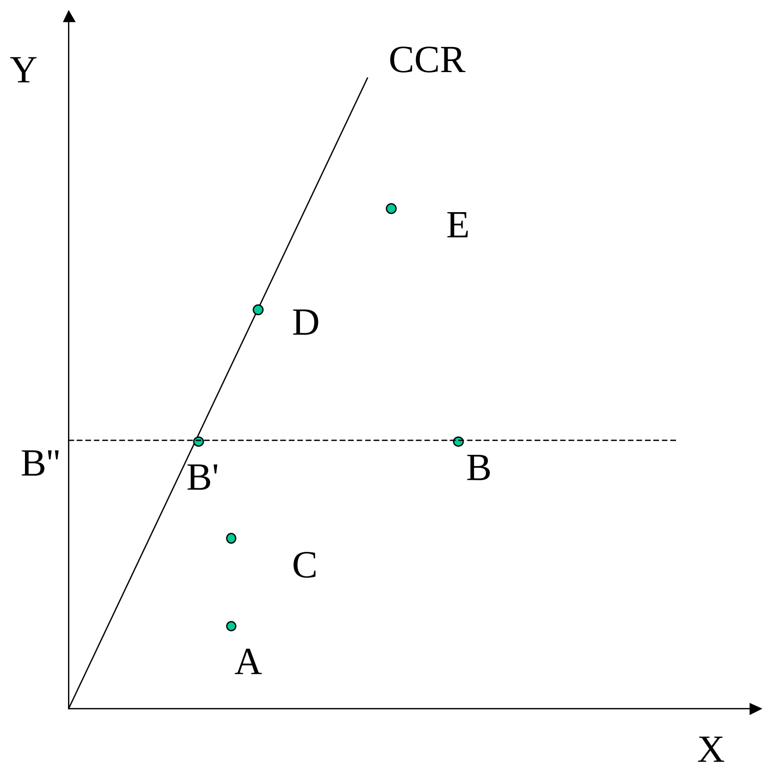

Figure 2.8 CCR Production Function

Figure 2.8 shows the CCR production function for the simple one-input (X) one-output (Y) case. Due to the constant returns to scale assumption this model gives identical results for both input-reducing and output-increasing orientations. Consider five DMUs, A, B, C, D, and E. Model M3 is solved and efficiency scores are calculated for each DMU. Physically these scores represent the excess input usage or shortfall in output production. Geometrically, the scores are a distance measure between each DMU and its horizontal projection (input orientation) or vertical projection (output orientation) onto the CCR production function. From Figure 2.8 the input-reducing efficiency score for DMU B is :

$$\mathrm{TE}_{\mathrm{Input}}(\mathrm{B}) = \frac{\mathrm{B''B'}}{\mathrm{B''B}} \tag{2.25}$$

With appropriate modifications, the CCR model provides the decision-maker with input and output target values that would transform inefficient units as efficient. The constructed virtual unit then represents targets for $\mathrm{DMU_c}$, the attainment of which would make the unit efficient. To obtain these target values the CCR model (M3) has to be rewritten as:

CCR Model with Slacks (M4)

$$Min \quad \theta_{\mathrm{c}} - \varepsilon\left(\sum_{i=1}^{m} e_i + \sum_{j=1}^{s} r_j\right) \tag{2.26}$$

$$\text{subject to} \quad \sum_{\mathrm{k=1}}^{\mathrm{n}} \mathrm{z_k} x_{ik} + e_i = \theta_c x_{ic}, \qquad i = \{1,2,\ldots,m\} \tag{2.27}$$

$$\sum_{k=1}^{n} z_k y_{jk} - r_j = y_{jc}, \qquad j = \{1,2,\ldots,s\} \tag{2.28}$$

$$\theta_c, z_k, e_i, r_j \geq 0, \qquad \forall\ i, j, k \tag{2.29}$$

where e_i and r_j are the slack variables introduced to convert the constraints from inequalities to equalities. DMU_c is efficient when the slacks are equal to zero. When DMU_c is inefficient then the input-efficiency score $\theta_c^* \leq 1$ and/or $(e_i, r_j) > 0$.

2.3.2 The BCC Model

In DEA an inefficient DMU can be made efficient by projection onto the efficient frontier or the envelopment surface. However, the DEA model used determines the actual point of projection that is chosen on the envelopment surface (Charnes *et al.* (1993)).

The CCR model assumes constant returns to scale *i.e.*, if all inputs are increased proportionally by a certain amount then the outputs will also increase proportionally by the same amount. However, Banker, Charnes, and Cooper (1984) noted that the constant returns to scale assumption skewed the results when making comparisons among DMUs differing significantly in size. In such situations it would be pertinent to know how the scale of operation of a DMU impacts its (in)efficiency. Thus, Banker *et al.* (1984) developed a new formulation of data envelopment analysis that is commonly known as the BCC model. The BCC model enables the use of a new empirical production function and is used to compute efficiency under the assumption of variable returns to scale *i.e.*, a proportional increase in inputs need not necessarily yield a proportional increase in outputs.

Whereas the CCR model addresses aggregate (technical and scale) efficiency, the BCC model addresses pure technical and *scale* efficiency. Efficiency is made up of technical (physical) efficiency and scale efficiency. Scale efficiency is explained through Figure 2.9 (Boussofiane *et al.* (1991)). DMU C is inefficient as it is enveloped by the efficient frontiers. The input-reducing technical efficiency for C at its scale of operation is given as X_A/X_C (C must reduce its inputs to A's level to become efficient as both produce the same output). B' is both technical and scale efficient or *aggregate* efficient. Further, it is the most aggregate efficient unit in the production possibility set. C's aggregate efficiency can be calculated by comparing it with B' or B (since B and B' lie on the same line they have the same slope and therefore the same numerical productivity). Therefore, the aggregate efficiency of C is given as:

$$\frac{OX_B}{OX_C} = \frac{OX_B}{OX_A} * \frac{OX_A}{OX_C} \tag{2.30}$$

i.e., Aggregate Efficiency = Scale Efficiency * Technical Efficiency

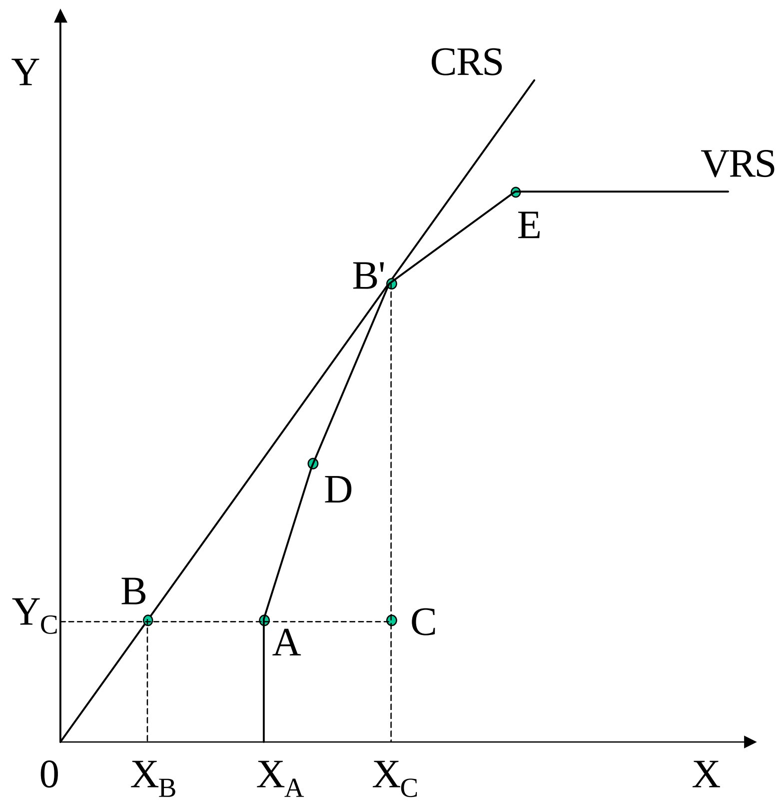

Figure 2.9 Technical Efficiency and Scale Efficiency

The output-increasing technical and scale efficiencies can be computed in an analogous manner. Thus, the BCC model enables measurement of both technical and scale efficiencies. An additional convexity constraint appears in the BCC model. This constraint restricts the sum of the activity levels of the input and output factors to one and restricts the virtual DMU to be of the same scale size as the DMU under consideration.

The BCC model focuses on maximal movement toward the frontier by proportional reduction of inputs (input-reducing) or by proportional augmentation of outputs (output increasing) (Charnes *et. al.,* 1994). The mathematical representation of the two-stage BCC model is given in the next section.

2.3.2.1 BCC: Input-Orientation (M5)

$$Min \quad z_0 = \theta - \varepsilon \sum_{j=1}^{s} s_j^+ - \varepsilon \sum_{i=1}^{m} s_i^- \tag{2.31}$$

$$subject\ to \quad \theta x_{ic} - s_i^- = \sum_{k=1}^{n} x_{ik} z_k, \qquad i = \{1,2,\ldots,m\} \tag{2.32}$$

$$\sum_{k=1}^{n} y_{jk} z_k - s_j^+ = y_{jc}, \qquad j = \{1,2,\ldots,s\} \tag{2.33}$$

$$\sum_{k=1}^{n} z_k = 1 \tag{2.34}$$

$$z_k, s_i^-, s_j^+ \geq 0 \tag{2.35}$$

In the formulation above, the objective function contains both the variable θ and the non-Archimedean (infinitesimally small) constant ε. Equation (2.34) represents the additional convexity constraint. The dual of this formulation would show that ε acts as a lower bound for the dual multipliers. The scalar variable θ is the proportional reduction of all inputs for the DMU under consideration which would then improve its efficiency. The simultaneous reduction of all inputs causes a radial movement toward the envelopment surface. Therefore, a DMU is efficient if and only if (i) $\theta^* = 1$, and (ii) all slacks are zero. The radial efficiency measure (input orientation), thus computed by the BCC model can be arrived at via a two-stage process *i.e.*, first, the maximal reduction in inputs given by θ* is calculated. This however, does not guarantee that the DMU will move onto the *efficient subset* through the equi-proportional reduction in inputs. Therefore, the second stage helps to determine the input surplus e^+ and the output slack r^-. Decision-makers can thus identify causes and quantities of inefficiencies through non-zero-slacks and a θ* value less than 1.

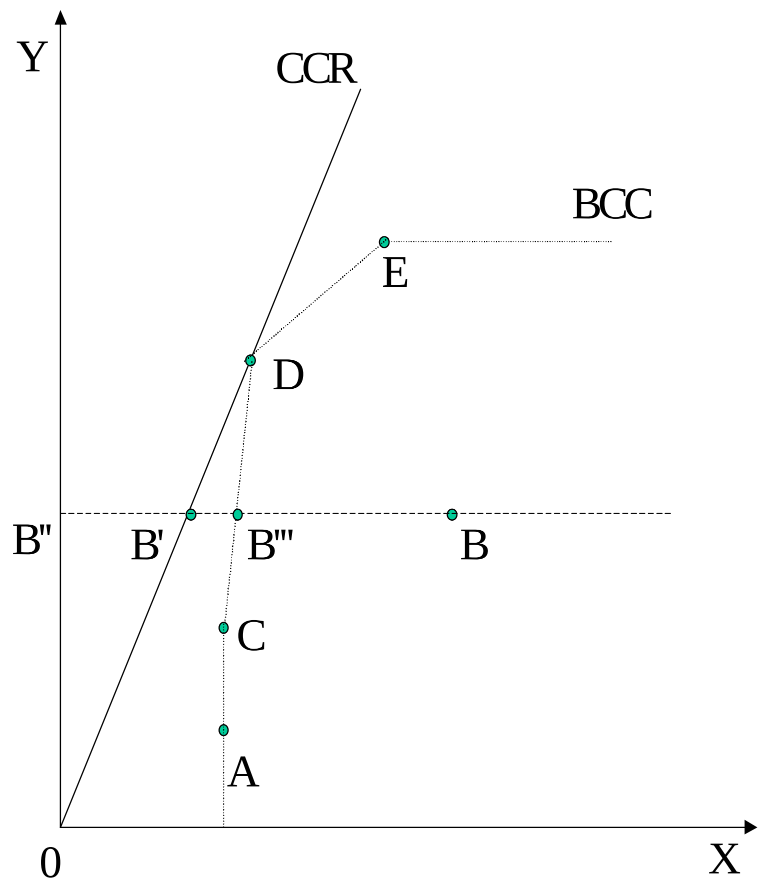

Figure 2.10 CCR and BCC Production Functions

Radial measures, however, have limitations as stated by Färe and Lovell (1978). Firstly, such measures compare DMUs to the efficient frontier or isoquant and not the efficient subset of the isoquant. This sometime results in a DMU using excess inputs also being termed efficient as compared to a DMU on the efficient subset. Secondly, the radial measure of technical efficiency is essentially based on Farrell's (1957) assumptions of the production function which limits its application to production technologies that satisfy those assumptions. And lastly, radial measures involve proportional reduction/augmentation of input/output mixes respectively which is not always feasible in real world scenarios. Färe and Lovell (1978) developed a non-radial measure that addresses these shortcomings for the production function by terming only DMUs on the efficient subset as efficient and by scaling input factors by different proportions to define the path of projection onto the efficient subset. The Färe-Lovell input-reducing technical efficiency measure is presented in the next section.

2.3.3 Färe-Lovell Input-Reducing Technical Efficiency Model

Consider a production technology that transforms inputs $x = (x_1, x_2,\ldots, x_m)$ into outputs $y = (y_1, y_2,\ldots, y_s)$ and let λ_i ($i = 1,2,\ldots, m$) be scalar weights associated with inputs x_i. Then Färe and Lovell (1978) define the Russell measure of input efficiency as:

$$R\ (x, y) = \mathrm{Min}\ \{\Sigma\ \lambda_i/m\text{: } \lambda_i x_i \in \mathrm{L}(y),\ \lambda_i \in (0,1]\ \forall\ i\}, \tag{2.36}$$

where $\mathrm{L}(y)$ is defined as in section 2.1.1.

The scalar weight λ_i is the contraction in each input i. The Russell measure minimizes the average contraction over all the inputs. The point of projection on the efficient subset is obtained by reducing each input by different proportions or by λ_i (Figure 2.11).

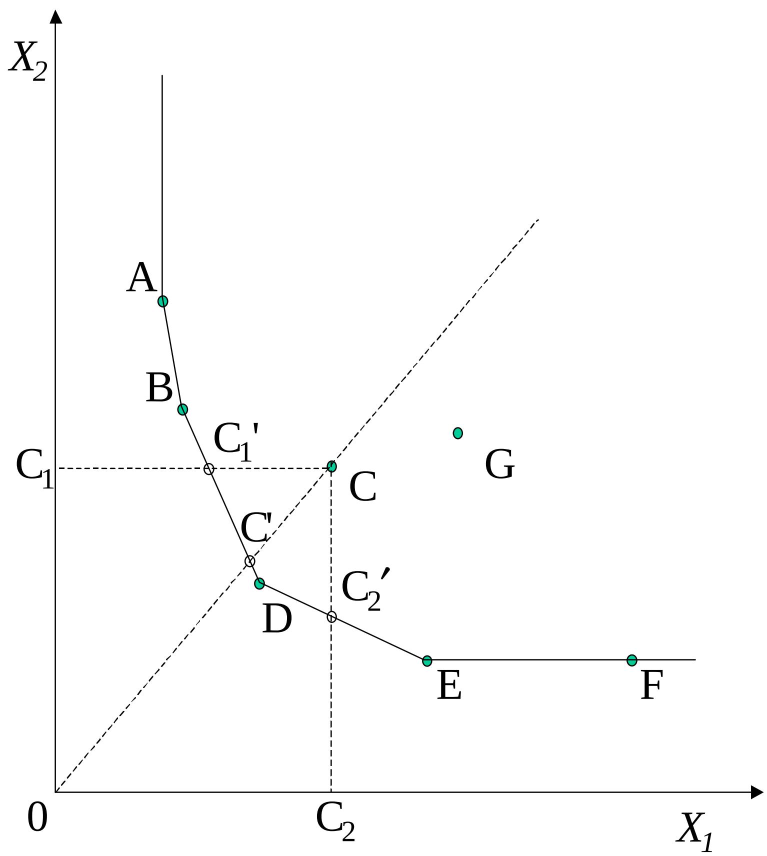

Figure 2.11 Färe-Lovell Input Reducing Technical Efficiency Measure

Figure 2.11 shows DMUs A, B, C, D, E, F, and G consuming inputs X_1 and X_2 to produce a certain output Y. C is an inefficient unit. The Farrell (1957) measure of technical efficiency would radially project C onto unit C′. The Russell measure would project C onto either D or E by reducing the inputs in varying proportions to reach the efficient subset. However, the minimization of the average reduction in inputs would choose one of D or E depending on the numerical outcome of equation (2.36) above.

The Färe-Lovell non-radial input-reducing technical efficiency measure is formulated mathematically as:

Färe-Lovell Non-Radial Input-Reducing Technical Efficiency Measure

$$Min\ \frac{1}{m}\sum_{i=1}^{m}\lambda_i \qquad (2.37)$$

$$\text{subject to} \quad \sum_{k=1}^{n} z_k x_{ik} \le \lambda_i x_{ik}, \qquad i = \{1,2,\ldots,m\} \qquad (2.38)$$

$$\sum_{k=1}^{n} z_k y_{jk} \ge y_{jc}, \qquad j = \{1,2,\ldots,s\} \qquad (2.39)$$

$$\sum_{k=1}^{n} z_k = 1 \qquad (2.40)$$

$$\lambda_i \le 1, \qquad i = \{1,2,\ldots,m\} \qquad (2.41)$$

$$\lambda_i,\ z_k \ge 0, \qquad \forall\ i,k \qquad (2.42)$$

2.3.4 DEA Applications

Accountability for the social and economic performance of public sector organizations has been a growing concern of society (Lewin and Morey (1981)). Thanassoulis (1996) adopts a DEA approach to contrast schools based on their differential effectiveness on pupils through the grade distribution while considering factors such as the pupils' family background, abilities, and overall effectiveness of the school. Athanassopoulos (1997) uses DEA along with managerial value judgements to assess the productive efficiency of Greek bank branches through (i) the operating efficiency of the branch and (ii) the quality of the service provided by the branch. Athanassopoulos and Ballantine (1995) use DEA to complement ratio analysis in measuring corporate performance for the grocery industry in the UK. The assessment includes sales' efficiency, effect of economies of scale, and performance benchmarking. Färe and Primont (1993) model the hierarchical structure of multi-unit banking via DEA to provide insights on the possible benefits of consolidation of banks. Bookbinder (1993) uses DEA to compare the performance of North American railroads. Viitala and Hanniner (1998) study the efficiency of public forestry organizations through DEA with Tobit models. An extensive reference list of DEA applications can be found in Seiford (1996).

2.4 MULTI-LEVEL PROGRAMMING

Many real life systems are characterized by hierarchical structures. The problem of resource allocation in and coordination of such systems is very complex. Multi-unit multi-level programming in organizations with a hierarchical structure has received extensive attention in the literature. A detailed review is provided by Nijkamp and Rietveld (1981), Burton and Obel (1977), Sweeny *et al.* (1978), Ruefli (1974), and Nachane (1984). A special case of the multi-level programming (MLP) problem is the linear bilevel programming problem (BLP). Wen and Hsu (1991) provide a review of the BLP with the basic models, characterizations, solution approaches, and application areas.

Multi-level programming is a mathematical programming approach used to solve decentralized planning problems. The solution procedure for multi-level programming problems involves the decomposition of the global problem into a number of smaller independent problems. In a multi-level programming problem the system is composed of an upper level or superordinate and one or more lower levels or subordinates. A decision-maker at any hierarchical level may have his own objective function and constraints and may also be influenced by other levels or other units at the same hierarchical level. The relationships between the different levels and within levels may be implicit or explicit. It is assumed that the preferences of the decision-makers over objectives at different levels may diverge and often be conflicting.

According to Anandalingam (1988) the problem in a multi-level system is to achieve the overall organizational goals while all decision-makers at all levels try to satisfy their own goals. For example, a problem generally facing hierarchical organizations is the allocation of scarce resources. The multi-level planning problem is then to find a feasible solution to a set of objectives constrained by the available resources that will also contribute maximally to the plan or objectives developed by the superordinate. Multi-level programming provides a specific structure necessary to address planning problems in hierarchical organizations. Policy making in multi-level systems is characterized by three main problems as described by Nijkamp and Rietveld (1981). These problems are (i) interdependencies between the subsystems; (ii) conflicts between the goals, priorities, and targets within each subsystem; and (iii) conflicts between the goals, priorities and targets between subsystems. Multi-level programming is an approach proposed in the literature to address these issues. The global objectives that may conflict with the goals of various organizational levels can be compromised using multi-level programming and the conflicts between the organizational levels can be resolved through coordinating mechanisms. Several forms of coordinating mechanisms have been reported in the literature (Nachane (1984), Freeland and Baker (1975)). Coordinating mechanisms for solving multi-level models have been categorized into two main groups with varying terminology. Ruefli (1974) terms them as classical and behavioral models, Sweeney *et al.* (1978) term them as decomposition and composition approaches and Burton and Obel (1977) term them as the pricing and budgeting approaches.

The multi-level planning problem can be described conceptually as follows.

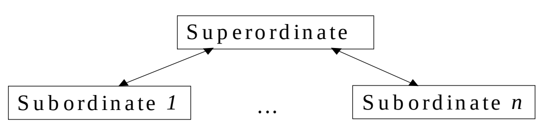

Figure 2.12 A Multi-Level System

The above system is composed of a supeordinate (upper level) and several subordinates (lower level), each of which has a decision-maker. This hierarchy of levels can extend through to several levels. The interrelationships in such systems where every decision-making unit at each level has its own objective function and set of constraints are very complex. Multi-level planning problems are thus very complex optimization problems that require computationally demanding algorithms. The mathematical representation of the MLP is given as follows (Bard (1984)).

$$\max f^{p}(x), \tag{2.43}$$

$$\text{where } x^{p-1} \text{ solves} \tag{2.44}$$

$$\max f^{p-1}(x), \tag{2.45}$$

$$\text{where } x^{p-2} \text{ solves} \tag{2.46}$$

$$\vdots$$

$$\max f^{1}(x), \tag{2.47}$$

$$\text{s.t. } x \in S. \tag{2.48}$$

In the above formulation f^i is the objective function of the i^{th} level ($i = 1,2,\ldots,p$) in the hierarchy defined over a jointly dependent feasible set S and is to be maximized by the respective decision-makers. An assumption is made that the decisions are made sequentially beginning with decision-maker p who controls decision vector $x^p \in X^p$, followed by decision-maker p-1 who controls decision vector $x^{p-1} \in X^{p-1}$ and so on through decision-maker 1 who controls decision vector $x^1 \in X^1$.

The multi-level programming approach has its roots in the decomposition method of Dantzig and Wolfe (1961). However, the solution process of the decomposition method is analogous to solving a single objective optimization problem over a fixed feasible region where the assumption is made that the aggregate objectives of the subordinates is the objective of the superordinate. Kornai and Liptak (1965) first introduced multi-level planning in the context of resource allocation in a scenario involving a central planning agency with independent sectors working under it. Candler and Townsley (1982) advanced a formal definition for the MLP problem. The multi-level programming problem adopts a game theoretic approach where each decision-making entity has a unique objective function and its own set of constraints. Most of the work in multi-level programming has been focused in developing solution approaches to the bilevel programming (BLP) problem. The linear bilevel programming problem is a special case of the MLP problem with a two-level hierarchical structure. The mathematical formulation of the BLP problem is given below.

$$\text{Max } F(x,y) = ax + by \tag{2.49}$$

$$\text{where } y \text{ solves} \tag{2.50}$$

$$\max f(x,y) = cx + dy \tag{2.51}$$

$$\text{s.t. } Ax + By \leq r \tag{2.52}$$

where $a, c \in \mathbf{R}^{n1}$ $b, d \in \mathbf{R}^{n2}$ $r \in \mathbf{R}^{m}$, A is an m x n_1 matrix, B is an m x n_2 matrix, x and y decision variables of the upper and lower problems respectively; F and f are the objective functions of the upper and lower problems respectively. This formulation captures the organizational hierarchy in which the decision-makers at upper and lower levels have to better their decisions from a jointly dependent constraint region S = $\{(x,y): Ax + By \leq r;\ x,y \geq 0\}$. The superordinate makes his decision first and thus fixes x. The subordinate decision-maker then uses the resultant solution from the superordinate (say x_0) and then solves the inner (lower level) problem to find y. The BLP problem is open to many interpretations regarding the variables controlled by the different levels, the autonomy of the subordinates, the influence of the superordinate and the restrictions on the decision vectors. However, the above formulation of the BLP problem is reported in the literature as the general formulation.

Wen and Hsu (1991) generalize the solution approaches for the linear BLP problem in two categories. The solution approaches in the first category search for the optimal solution among the extreme points of the original convex polyhedron (Candler and Townsley (1982), Bialas and Karwan (1984), Bard (1983), (1984)). The solution approaches in the second category use the corresponding Kuhn-Tucker conditions in place of the linear BLP problem and then require a solution to a set of equalities and inequalities (Fortuny-Amat and McCarl (1981), Bard and Falk (1982), Bialas and Karwan (1984), Wen and Bialas (1986)).

2.5 GOAL PROGRAMMING

Decision-making in a hierarchical organization is characterized by multiple goals or objectives at different levels that are often conflicting and incommensurable (measured in different units). Goal programming (GP) is a mathematical programming approach that incorporates various goals or objectives which cannot be reduced to a single dimension. Charnes and Cooper (1961) originated GP to solve goal-resource problems, which when modeled with linear programming techniques were found to have infeasible solutions. GP is also a good approach to solve multi-criteria decision making problems with conflicting objectives (Charnes and Cooper (1961), Ignizio (1976)). GP serves as a good decision tool in modeling real world problems involving multiple objectives.

The fundamental concept in GP is to incorporate all goals of the decision-maker in the model formulation. GP can handle single or multiple goals. In reality, the decision-maker generally chooses the achievement of certain goals at the expense of others. Therefore, GP requires an ordinal ranking of the goals in order of importance by the decision-maker. The solution process then satisfies goals beginning with the goal with the highest priority. Lower order goals are considered only after higher order goal shave been satisfied. The solutions to the higher order goals then become constraints for the lower order goals. As such it may not be possible to satisfy all goals to the required extent. GP can then be used to find a satisfactory level of achievement of the goals. Accordingly it is necessary to specify aspiration levels for the goals. Therefore, the objective function deals with minimization of the positive and negative deviations from the goals following the preemptive priorities assigned to the deviations.

The general formulation of a goal programming model is as follows.

$$\min Z = \Sigma\, w_i(d_i^+ + d_i^-) \tag{2.53}$$

$$\text{s.t. } \Sigma\, a_{ij}x_j + d_i^- - d_i^+ = b_i \qquad \forall\, I \tag{2.54}$$

$$X_{j,}\ d_i^-, d_i^+ \geq 0 \qquad \forall\, i, j \tag{2.55}$$

where x_j represents a decision variable, w_i represents the weights attached to goal i, and d_i^-,and d_i^+ represent the under achievement and over achievement of a goal i respectively.

GP, however, due to the nature of its objective function sometimes tends to overachieve certain goals while underachieving others. Goal interval programming (Charnes and Cooper (1977)) addresses this shortcoming by specifying an interval within which all points are equally desirable towards achievement of the target goal.

2.6 GOAL PROGRAMMING AND DATA ENVELOPMENT ANALYSIS (GODEA)

Athanassopoulos (1995) developed a model integrating Goal Programming and Data Envelopment Analysis (GoDEA) to incorporate target setting and resource allocation in multi-level planning problems. The GoDEA framework is proposed as a decision-making tool that combines conflicting objectives of efficiency, effectiveness and equity in resource allocation as well as incorporates viewpoints of different management levels in the planning process. Thanassoulis' and Dyson's (1992) formulation provides a method to estimate input/output targets for each individual DMU in a system but fails to address planning and resource allocation issues at the global organizational level while considering all DMUs simultaneously. That is, their formulation does not carry global organizational targets or global resource constraints. Athanassopoulos (1995) provided these enhancements with his formulation of the GoDEA model. The mathematical representation of his model is given as:

GoDEA Model

$$\underset{p_i, p_j, n_i, n_j, d_i, d_j, \lambda_k}{Min} \; \{\sum_{k=1}^{N}\sum_{i \in Iq}(P_i^n \frac{n_i^k}{x_{ik}} + P_i^p \frac{p_i^k}{x_{ik}}) + \sum_{k=1}^{N}\sum_{j \in J_q}(P_j^n \frac{n_j^k}{y_{jk}} + P_j^p \frac{p_j^k}{y_{jk}}), \quad (2.56)$$

$$\sum_{i \in I_V} P_i^g \frac{d_i^+}{GX_i} + \sum_{j \in J_V} P_j^g \frac{d_j^-}{GY_j}\}$$

Subject to:

DMU representation:

$$\sum_{k=1}^{N} \lambda_k^c y_{jk} - p_j^c + n_j^c = y_j^c, \qquad j \in J, \;\; \forall c \qquad (2.57)$$

$$-\sum_{k=1}^{N}\lambda_k^c x_{ik} + p_i^c - n_i^c = -x_i^c, \qquad i \in I, \quad \forall c \tag{2.58}$$

Effectiveness through Achievement of Global Targets:

$$\sum_{k=1}^{N}\lambda_k^1 y_{jk} + \ldots + \sum_{k=1}^{N}\lambda_k^N y_{jk} + d_j^- = GY_j, \qquad \forall j \in J \tag{2.59}$$

$$-\sum_{k=1}^{N}\lambda_k^1 x_{ik} + \ldots + \sum_{k=1}^{N}\lambda_k^N x_{ik} + d_i^+ = -GX_i, \qquad \forall i \in I \tag{2.60}$$

Budget Balance:

$$\sum_{i \in I_B}\sum_{k=1}^{N}(\lambda_k^1 + \ldots + \lambda_k^N)x_{ik} - \sum_{j \in J_B}\sum_{k=1}^{N}(\lambda_k^1 + \ldots + \lambda_k^N)y_{jk} \leq \mathrm{B}, \forall i \in I_B \text{ and } j \in J_B \tag{2.61}$$

$$\lambda_k^c, n_i^k, n_j^k, p_i^k, p_j^k, d_i, d_j \geq 0 \tag{2.62}$$

where:

N:	number of DMUs,
I:	set of inputs,
J:	set of outputs,
x_{ik}:	level of input i for DMU k,
y_{jk}:	level of output j for DMU k,
x_i^c, y_j^c:	level of input i and output j for DMU c when assessing DMU c,
λ_k^c:	activity level of DMU k when assessing DMU c,
n_i^k, p_i^k:	negative and positive deviation variables for input i of DMU k,
n_j^k, p_j^k:	negative and positive deviation variables for output j of DMU k,
d_i^+, d_i^-:	positive and negative deviation variables from global targets of input i and output j,

P_i^n, P_i^p: user defined preferences over the minimization of positive and negative goal deviations of input *I*,

P_j^n, P_j^p: user defined preferences over the minimization of positive and negative goal deviations of output *j*,

P_i^g, P_j^g: user defined preferences related to global targets of input *i* and output *j*,

GX_i, GY_j: global target levels known *a priori* for input *i* and output *j*,

B: user specified constant for the budget balance constraint between commensurable inputs and outputs,

I_B, J_B: subsets of commensurable inputs ($I_B \subset I$) and outputs ($J_B \subset J$)

The GoDEA model outlined above is presented in condensed form as compared to Athanassopoulos' (1995) original formulation for reasons of simplicity. Athanassopoulos' (1995) formulation contains added complexity by way of the following dimensions. The set of inputs and the set of outputs are divided into subsets that are prioritized and not prioritized for improvement. The DEA type constraints contain deviation variables for the prioritized inputs/outputs and consequently these constraints are linked to the objective function. The DEA type constraints for the non-prioritized inputs/outputs contain no deviation variables. Further, the subsets of inputs and outputs prioritized to be improved are divided into controllable and uncontrollable inputs and outputs. The controllable inputs and outputs have associated global targets that are known *a priori* to the solution of the GoDEA model while the uncontrollable inputs and outputs have global targets that are estimated by the solution process of the model.

The complexity of Athanassopoulos'(1995) GoDEA formulation is not critical to the theoretical concepts proposed in this research. The simplified version of the model aims to facilitate presentation of the fundamental concepts proposed by Athanassopoulos (1995) and provide a basic framework for the reformulation of the GoDEA model as well as the fuzzy formulation proposed in this research and described in Chapter 3. The features of Athanassopoulos' (1995) GoDEA model in its simplified form are explained next.

The objective function seeks to minimize the deviation variables corresponding to the global targets and the individual DMU targets. The deviation variables are normalized in the objective function by division with the respective input/output to obtain a standard evaluation system *i.e.,* a system where all deviations have equal importance (weight) irrespective of the numerical value. The first part of the objective function contains the penalty per unit deviation from the global targets. It is assumed that the individual DMUs would always require a higher consumption of inputs than available and would produce a lower amount of outputs than required by higher management (target setting level). Therefore, only deviation variables corresponding to over achievement of global inputs and underachievement of global outputs are present in the objective function. This assumption can be appropriately relaxed by modifying the deviation variables. The second part of the objective function contains priorities for the negative and positive deviations that are used to track the contribution of individual DMUs to the global organizational targets. The essential difference between conventional DEA and Athanassopoulos' (1995) GoDEA model is the presence of two-way deviation variables which allow under and over achievement of inputs and outputs while DEA assumes that the inputs should always be contracted and outputs should always be expanded.

The first set of constraints captures the simultaneous representation of all DMUs within the planning process. These constraints compare the inputs/outputs of the assessed DMU with its composite unit. Goal deviation variables are used to allow for under and over achievement of the goals. The second set of constraints represents operational effectiveness. Here, the activities of all DMUs are aggregated and measured against global input/output targets that are allocated between or produced by the DMUs[2]. The degree of satisfaction of global targets gives a measure of operational effectiveness through the aggregate contribution of efficient DMU reference sets. Finally, the budget balance constraint provides one form of policy constraint. This constraint represents a balance relationship between aggregate input and output target achievements for

[2] The estimation of targets is not important in the context of this research. The reader is directed to Athanassopoulos (1995) for details on estimation of targets.

commensurable (measured in the same units) inputs and outputs (e.g., income-expenses relationship) when most efficient.

2.7 FUZZY DECISION MAKING

The realm of fuzzy decision making is discussed in this section. The literature review is divided into the following subsections. First, the mathematical concepts and notions of fuzzy set theory are introduced. Second, the linkage of fuzzy set theory and decision-making is established. In this subsection the definitions of fuzzy goal, fuzzy constraint, fuzzy decision, and optimal fuzzy decision are outlined. Third, the connection between fuzzy decision making and linear programming is discussed. In particular the model formulation proposed by Zimmermann (1976) and adapted to DEA by Sengupta (1992) is illustrated. Sengupta's (1992) formulation is adapted to the GoDEA model (Athanassopoulos (1995)) and developed in this research to provide a fuzzy decision-making environment incorporating goal programming and data envelopment analysis.

2.7.1 Fuzzy Set Theory

Fuzzy set theory made its first official appearance in Lofti Zadeh's (1965) famous paper titled "Fuzzy Sets" wherein he defined the fundamental postulates and introduced the theoretical basis. Zadeh (1965, p. 338) defined a fuzzy set as "*a class of objects with a continuum of grades of membership […] and characterized by a membership (characteristic) function which assigns to each object a grade of membership ranging between zero and one.*" Zadeh (1965) provided the following definitions related to fuzzy sets.

Suppose X is a space of objects and a generic element of X is denoted by x. Then the fuzzy set A in X is defined as the set of ordered pairs:

$$A = \{(x, \mu_A(x) | x \in X\} \tag{2.63}$$

The fuzzy set A in X is characterized by a membership function $\mu_A(x)$ such that each point in X is associated with a real number in the interval [0,1]. The value of $\mu_A(x)$ denotes the degree of membership of x in A and, therefore, the closer the value of $\mu_A(x)$ to unity the higher is degree of membership of x in A. When $\mu_A(x)$ takes on only two values 1 and 0 corresponding to whether x does or does not belong to A then $\mu_A(x)$ reduces to the ordinary characteristic function of A (*i.e.*, A is a non-fuzzy set).

Zadeh (1965) extended the definitions for ordinary sets to derive definitions for fuzzy sets. These definitions are consistent with topological concepts such as equality, complementation, containment, union, intersection, algebraic product, algebraic sum, normality, support, relation, composition, mapping, convexity, and concavity. These definitions are outlined below.

Empty Set: A fuzzy set A is empty if and only if its membership function is identically zero on X.

$$i.e., \mathrm{A} = \phi \Leftrightarrow \mu_A(x) = 0 \qquad \forall\, x \in X \tag{2.64}$$

Equality: Two fuzzy sets A and B are equal if and only if their membership functions are equal for all $x \in X$.

$$i.e., A = B \Leftrightarrow \mu_A(x) = \mu_B(x) \quad \forall\, x \in X \tag{2.65}$$

Complementation: The complement of a fuzzy set A is a fuzzy set A' with a membership function $\mu_{A'}(x)$ and is defined as

$$\mu_{A'}(x) = 1 - \mu_A(x) \qquad \forall\, x \in X \tag{2.66}$$

Containment: Fuzzy set A is contained in fuzzy set B (or A is a subset of B) if and only if $\mu_A(x) \le \mu_B(x)$ for all $x \in X$.

$$i.e., A \subset B \Leftrightarrow \mu_A(x) \le \mu_B(x) \quad \forall\, x \in X \tag{2.67}$$

Union: The union of two fuzzy sets A and B with membership functions $\mu_A(x)$ and $\mu_B(x)$ respectively is defined as a fuzzy set C with a membership function $\mu_C(x)$ such that C is the *smallest* fuzzy set containing both A and B.

$$\therefore\ \mu_C(x) = \mu_A(x) \vee \mu_B(x) \tag{2.68}$$

$$i.e., \quad \mu_C(x) = \mathrm{Max}\,[\mu_A(x), \mu_B(x)] = \mu_A(x) \quad \text{if } \mu_A(x) \geq \mu_B(x) \tag{2.69}$$

$$\mu_C(x) = \mathrm{Max}\,[\mu_A(x), \mu_B(x)] = \mu_B(x) \quad \text{if } \mu_A(x) \leq \mu_B(x) \tag{2.70}$$

Note: $\vee$ has the associative property, i.e., $A \vee (B \vee C) = (A \vee B) \vee C$

Intersection: The intersection of two fuzzy sets A and B with membership functions $\mu_A(x)$ and $\mu_B(x)$ respectively is defined as a fuzzy set C with a membership function $\mu_C(x)$ such that C is the *largest* fuzzy set contained in both A and B.

$$\therefore\ \mu_C(x) = \mu_A(x) \wedge \mu_B(x) \tag{2.71}$$

$$i.e., \quad \mu_C(x) = \mathrm{Min}\,[\mu_A(x), \mu_B(x)] = \mu_A(x) \quad \text{if } \mu_A(x) \leq \mu_B(x) \tag{2.72}$$

$$\mu_C(x) = \mathrm{Min}\,[\mu_A(x), \mu_B(x)] = \mu_B(x) \quad \text{if } \mu_A(x) \geq \mu_B(x) \tag{2.73}$$

Algebraic Product: The algebraic product of fuzzy sets A and B is denoted as AB and defined such that for all $x \in X$:

$$\mu_{AB}(x) = \mu_A(x)\mu_B(x) \tag{2.74}$$

Algebraic Sum: The algebraic sum of fuzzy sets A and B is denoted as $A \oplus B$ and defined such that for all $x \in X$:

$$\mu_{A \oplus B}(x) = \mu_A(x) + \mu_B(x) - (\mu_A(x)\mu_B(x)) \tag{2.75}$$

Relation: A fuzzy relation R in the product space X_1 x X_2 is a fuzzy set with a membership function $\mu_R(x) : X_1 \text{ x } X_2 \quad R$ which associates a degree of membership $\mu_R(x_1, x_2)$ in R with each ordered pair (x_1, x_2).

Decomposition: Consider fuzzy set C in X x Y = {x, y} with a membership function $\mu_C(x,y)$ and fuzzy sets A and B with membership functions $\mu_A(x)$ and $\mu_B(y)$ respectively. Then C is decomposable along X and Y if and only if:

$$\mu_C(x) = \text{Min}\ (\mu_A(x), \mu_B(y)) \tag{2.76}$$

Mapping: Consider T: $X \rightarrow Y$ a mapping from X to Y. Let B be a fuzzy set in Y with a membership function $\mu_B(y)$. The inverse mapping T^{-1} induces a fuzzy set A in X with a membership function

$$\mu_A(x) = \mu_B(y), \qquad y \in Y \tag{2.77}$$

for all x $\in$ X which are mapped by T into Y.

Now, consider conversely that A is a fuzzy set in X. Then T induces a fuzzy set B in Y such that:

$$\mu_B(y) = \text{Max}_{x \in T^{-1}(y)}\ (\mu_A(x)), \qquad y \in Y \tag{2.78}$$

where $T^{-1}(y)$ is the set of points in X which are mapped into Y by T.

Concavity and Convexity: A fuzzy set A' is concave if its complement A is *convex*. A fuzzy set A is convex if and only if for every x_1, $x_2 \in X$ and all $\beta \in [0, 1]$,

$$\mu_A(\beta x_1 + (1-\beta)x_2) \geq \text{Min}\ (\mu_A(x_1), \mu_A(x_2)) \tag{2.79}$$

Normality: A fuzzy set A is normal if and only if the supremum of $\mu_A(x)$over X, $\text{Sup}_x\ \mu_A(x)$, is equal to 1. Otherwise A is subnormal.

Support: The support of a fuzzy set A is a subset of A, S(A), such that:

$$x \in S(A) \Leftrightarrow \mu_A(x) > 0. \tag{2.80}$$

2.7.2 Decision Making In A Fuzzy Environment

Bellman and Zadeh (1970) extended fuzzy set theory and developed a framework for decision-making in a fuzzy environment. A fuzzy environment is defined as an environment where the goals and/or constraints are fuzzy. Conventional decision-making environments consist of three parts, namely, objectives, constraints and alternatives. The alternatives define the decision space (from which a solution may be chosen) and is restricted by the constraints. The objective(s) or goal provides the selection criteria for the solution and assigns a utility value to all possible choices.

Bellman and Zadeh (1970) forwarded the premise that objectives and constraints can be treated as fuzzy sets in the decision space and a fuzzy decision then would be obtained as the intersection of these fuzzy sets. Their conceptual definition is stated as:

"Decision = Confluence of Goals and Constraints" (2.81)

Accordingly, the formal definition (Bellman and Zadeh (1970) is stated as:

> *"Assume that we are given a fuzzy goal G and a fuzzy constraint C in a space of alternatives X. Then G an C combine to form a decision D, which is a fuzzy set resulting from intersection of G and C. In symbols, $D = G \cup C$ and correspondingly, $\mu_D(x) = \mu_{G \cup C}(x)$."*

In the general case, suppose there are n goals G_i (i = 1, 2,…,n) and m constraints C_j (j = 1, 2,…,m). The resultant fuzzy decision is given as the intersection of the n goals and m constraints as:

$$D = G_1 \cup G_2 \ldots \cup G_n \cup C_1 \cup C_2 \quad \cup C_m \tag{2.82}$$

$$\text{i.e., , } \mu_D(x) = \text{Min } (\mu_{G1}(x),\ldots, \mu_{Gn}(x), \mu_{C1}(x),\ldots, \mu_{Cm}(x)) \tag{2.83}$$

Given a set of fuzzy decisions Bellman and Zadeh (1970) addressed the optimal fuzzy decision. They proposed that the optimal fuzzy decision in the decision space *X* was a maximizing decision *i.e.,* a decision that maximized $\mu_D(x)$.

2.7.3 Fuzzy Linear Programming

Dantzig introduced linear programming (LP) in the 1940s to model decision-making problems. LP involves maximization or minimization of a linear objective function subject to satisfaction of a set of linear constraints. The canonical form of LP is stated as (Bazaara *et al.*(1990)):

$$\text{Min } z = cx \tag{2.84}$$

$$\text{subject to: } \quad Ax \leq b \tag{2.85}$$

$$x \geq 0 \tag{2.86}$$

where $x = [x_j]$ is the vector of decision variables, $c = [c_j]$ is the vector of cost coefficients, $A = [a_{ij}]$ is a matrix of technological coefficients and $b = [b_i]$ is the resource vector. A best possible solution from the decision space is obtained by solving the above linear program.

Fuzzy linear programming is used to model the decision-making environment with imprecise information regarding either all or some of the parameters *A*, *b*, and *c* where information is available regarding the interval range of the parameters. Further, the decision-maker may allow violation of the constraints within certain tolerance limits and/or require considerable improvement in the objective function rather than cost minimization. The solution to the fuzzy LP will yield a near optimal solution that will satisfy the constraints within the specified parameter ranges and/or constraint tolerance limits. This approach differs from the sensitivity analysis associated with conventional

LP. Sensitivity analysis is a post-optimization technique used to evaluate other optimal solutions in the neighborhood of the originally obtained optimal solution. In fuzzy LP the value of the parameters *A*, *b*, and *c* depend on the value of the membership function specified. Also, in the case of imprecise constraints, the optimal solutions depend on the membership functions associated with the constraints as well as the specified tolerance limits. Thus, different optimal solutions are obtained for different values of the membership function which need not necessarily be in the same neighborhood.

Zimmermann (1985) developed a fuzzy LP formulation for fuzzy objective function and constraints based directly on Bellman's and Zadeh's (1970) definition of fuzzy decision-making. Carlsson and Korhonen (1986) formulated a fuzzy LP to incorporate imprecise parameter values with knowledge of their upper and lower bounds. Their formulation is termed as *fuzzy parametric programming*.

2.7.4 The Zimmermann Model

Suppose that the decision-maker has target values for the objective function and tolerance limits for the constraints. Then, Zimmermann (1976) suggested a fuzzy formulation of the conventional LP as follows:

$$\text{Find } x \in X \tag{2.87}$$

$$\text{s.t.} \quad (cx)\ \delta\ (z) \tag{2.88}$$

$$(Ax)\ \delta\ (b) \tag{2.89}$$

$$x \geq 0 \tag{2.90}$$

where δ signifies *almost less than or equal to.*

The above model can be rewritten with matrix $B = (c, A)$ and vector $d = (z, b)$ as:

$$(Bx)\ \delta\ (d) \tag{2.91}$$

$$x \geq 0 \tag{2.92}$$

where each row of constraint $(Bx)\ \delta\ (d)$ is a fuzzy set with membership function $\mu_i(x)$ that represents the degree to which the constraint is satisfied by x_i. It should be noted that the objective function is also treated as a constraint in a fuzzy environment. The membership functions are determined subjectively but must mathematically satisfy the following relationships.

$$\mu_i(x) = 1 \qquad \text{if } Bx_i \leq di, \qquad i = \{1, 2, \ldots, \text{m+1}\} \tag{2.93}$$

$$\mu_i(x) \in [0, 1] \qquad \text{if } d_i \leq Bx_i \leq d_i + p_I, \qquad i = \{1, 2, \ldots, \text{m+1}\} \tag{2.94}$$

$$\mu_i(x) = 0 \qquad \text{if } Bx_i \geq d_i, \qquad i = \{1, 2, \ldots, \text{m+1}\} \tag{2.95}$$

where p_i are the subjectively chosen tolerance limits to represent acceptable violation of the constraints.

The optimal solution to the above problem is given by Zimmermann (1976) as the intersection of the $(m + 1)$ constraints. Therefore, the membership function associated with the fuzzy decision is:

$$\mu_D(x) = \text{Min}_{i = \{1, 2, \ldots, m+1\}} (\mu_i(x)) \tag{2.96}$$

Let $\lambda = \mu_D(x)$. Zimmermann (1976) transformed the above fuzzy LP to:

$$\text{Max } \lambda \tag{2.97}$$

$$\text{Subject to} \qquad \lambda \leq \mu_i(x) \tag{2.98}$$

$$0 \leq \lambda \leq 1 \tag{2.99}$$

$$x \geq 0 \tag{2.100}$$

where λ and x are unknown, the p_i's are the specified tolerance limits within which the constraints can be violated, and B and d are known crisp coefficients.

2.7.5 The Carlsson and Korhonen Model

Carlsson and Korhonen (1986) proposed to fuzzify the parameters of a conventional linear program while keeping the constraints crisp. Consider a conventional linear programming (p. 48) where the parameters A, b and c are not known precisely but the decision-maker is able to specify their lower and upper bounds. In such a case, the lower bounds can be interpreted as *risk-free values* (pessimistic) which are most certainly achievable and the upper bounds can represent *impossible values* (optimistic) which are unrealistic and cannot be attained. Conceptually, the lower and upper bounds on the parameters signify that as the values of A, b and c move from their risk-free values to their impossible values, the solution to the linear program moves from a very high degree to a very low degree of plausible implementation.

Let A^0, b^0, c^0 represent the lower bounds and A^1, b^1, c^1 represent the upper bounds of the parameters A, b, and c respectively. According to Carlsson and Korhonen (1986) the subjective membership functions chosen to be associated with fuzzy sets A, B, and c should be *"monotonically decreasing functions of the parameters"* as the functions denote the risk level associated with the individual coefficients a_{ij}, b_i and c_j. Therefore, the membership functions should equal zero at the upper (impossible) bound and equal one at the lower (risk-free) bound. For example, membership functions for A, b, and c can be expressed in terms of their bounds as:

$$\mu_A(a_{ij}) = a_{ij} - a_{ij}^1/a_{ij}^0 - a_{ij}^1, \qquad a_{ij} \in [a_{ij}^0, a_{ij}^1], \qquad i = \{1,\ldots,m\}, j = \{1,\ldots,n\} \tag{2.101}$$

$$\mu_b(b_i) = b_i - b_i^1/b_i^0 - b_i^1, \qquad b_i \in [b_i^0, b_i^1], \qquad i = \{1,\ldots,m\} \tag{2.102}$$

$$\mu_c(c_j) = c_j - c_j^1/c_j^0 - c_j^1, \qquad c_j \in [c_j^0, c_j^1], \qquad j = \{1,\ldots,n\} \tag{2.103}$$

Carlsson and Korhonen (1986) used the concept of confluence proposed by Bellman and Zadeh (1970) to derive the membership function μ that would provide the solution to the above problem and is given as:

$$\mu = \mu_A(a_{ij}) \cup \mu_b(b_i) \cup \mu_c(c_j) = \text{Min}\ \{\mu_A(a_{ij}), \mu_b(b_i), \mu_c(c_j)\} \tag{2.104}$$

This implies that the exactitude of the fuzzy decision μ is equivalent to that of the riskiest of the parameters. Carlsson and Korhonen (1986) also note that the best value of the objective function at a fixed level of risk μ is achieved when $\mu = \mu_A = \mu_b = \mu_c$ *i.e.*, the objective function is maximized at a particular precision level when all parameters are evaluated at the same precision level. This characteristic of Carlsson and Korhonen (1986) formulation allows the decision-maker to rewrite the conventional linear programming model by expressing each of the parameters *A*, *b*, and *c* as a function of the respective lower and upper bounds and membership function. Therefore, a_{ij}, b_i and c_j can be written as:

$$a_{ij} = (a_{ij}^0 - a_{ij}^1)\ \mu_A + a_{ij}^1 \tag{2.105}$$

$$b_i = (b_i^0 - b_i^1)\ \mu_b + b_i^1 \tag{2.106}$$

$$c_j = (c_j^0 - c_j^1)\ \mu_c + c_j^1 \tag{2.107}$$

The LP can now be written as:

$$\text{Max } z = ((c_j^0 - c_j^1)\ \mu + c_j^1)x \tag{2.108}$$

$$\text{s.t.} \tag{2.109}$$

$$((A^0 - A^1)\ \mu + A^1)x \leq b_i = (b_i^0 - b_i^1)\ \mu + b_i^1 \tag{2.110}$$

$$x \geq 0 \tag{2.111}$$

The membership functions used in the above formulation are linear and hence the model can be solved as a linear parametric program. However, non-linear forms may be used for the membership functions. In such cases, non-linear programming techniques have to be used to solve the model.

2.7.6 Fuzzy Goal Programming (FGP)

Narsimhan (1980) was the first to integrate the concepts of fuzzy set theory and goal programming. Tiwari *et al.* (1986) provide an extensive review of the various facets of fuzzy goal programming that have been researched by Hannan (1981, 1982), Narsimhan (1981), Ignizio (1982), Tiwari *et al.* (1985, 1986), and Rubin and Narsimhan (1984).

Multi-criteria decision problems generally involve the resolution of multiple conflicting goals to achieve a *"satisfycing"* solution rather than maximization objectives given a suitable aspiration level for each objective. The generalized goal programming approach seeks to minimize the negative (under achievement) and positive (over achievement) deviations from the goal targets. However, in most real life situations the aspiration levels for some or all objectives typically have an imprecise nature. For example, the profit of a company should be around 2 million dollars. In other words, the aspiration levels need a linguistic interpretation such as very good, good, and moderately good. To capture such scenarios it is appropriate to model the objective(s) and constraints with a certain specified tolerance limit. All the fuzzy goals and fuzzy constraints can be considered as fuzzy criteria.

The central theme of FGP is that systems with ill-defined or imprecise characteristics are first modeled as fuzzy models (Sinha *et al.* (1988)). The fuzzy models are then formulated as crisp equivalent models which can be solved with existing decision-making methodologies.

Consider a fuzzy criterion $h(\mathbf{X})$ with an imprecise aspiration level $\boldsymbol{b}$ and denoted as $h(\mathbf{X})\ \tau\ \boldsymbol{b}$ where $\mathbf{X} \in \boldsymbol{R}^n$ is the decision variable vector. This criterion is interpreted as

"$h(\mathbf{X})$ should be *essentially* greater than or equal to $\boldsymbol{b}$". The "τ" signifies the fuzzification of the criterion. With these concepts consider the following system with k fuzzy criteria expressed as:

$$\text{Find } \mathbf{X} \in \boldsymbol{R}^n \geq 0 \tag{2.112}$$

$$\text{s.t.} \quad h_i(\mathbf{X})\ \tau\ b_i \qquad i = 1, 2, \ldots, k \tag{2.113}$$

The solution to the above fuzzy goal program can be obtained by solving its crisp equivalent. Fuzzy set theory is used to associate a fuzzy set h_i with each fuzzy criterion $h_i(\mathbf{X})$ and a membership function μ_i with each fuzzy set h_i. Bellman's and Zadeh's (1970) concept of confluence is used to find a fuzzy decision set D with membership function $\mu_D(\mathbf{X})$ by aggregating over all the fuzzy sets h_i with membership functions μ_i. Here, $\mu_D(\mathbf{X})$ serves as the objective function and $\mathbf{X} \in \boldsymbol{R}^n$ is the solution that maximizes $\mu_D(\mathbf{X})$. The problem then is to define the membership functions μ_i and the tolerance limits for the aspiration levels.

For the above FGP suppose $l_i \leq b_i$ is the lower bound for the constraint $h_i(\mathbf{X})\ \tau\ b_i$ such that $h_i(\mathbf{X}) \geq b_i$ implies exact satisfaction and $h_i(\mathbf{X}) \leq l_i$ implies no satisfaction. Consequently, let $\mu_i = 0$ imply no satisfaction and $\mu_i = 1$ imply full satisfaction. Then $\mu_i = (0,1)$ corresponds to varying degrees of satisfaction when $l_i \leq h_i(\mathbf{X}) \leq b_i$. The higher the degree of satisfaction the nearer the value of μ_i is to 1. Therefore, a fuzzy set h_i is associated with the fuzzy criteria $h_i(\mathbf{X})\ \tau\ b_i$ and a linear membership function μ_i is defined as:

$$\mu_i = 1 \qquad \text{for } h_i(\mathbf{X}) \geq b_i \tag{2.114}$$

$$\mu_i = h_i(\mathbf{X}) - p_i/b_i - p_i \qquad \text{for } l_i \leq h_i(\mathbf{X}) \leq b_i. \tag{2.115}$$

$$\mu_i = 0 \qquad \text{for } h_i(\mathbf{X}) \leq l_i \tag{2.116}$$

The use of non-linear membership functions can be used alternatively but would require the use of non-linear solution procedures.

As described in Bellman and Zadeh (1970) the optimal decision $\mu_D(\mathbf{X})$ is then given as:

$$\text{Max } \mu_D(\mathbf{X}) = \text{Max } \{\cup \mu_i\} \qquad (i = 1,\ldots,k) \tag{2.117}$$

$$= \text{Max } \lambda \text{ (say)} \tag{2.118}$$

$$= \text{Min } \{\mu_i\} \qquad (i = 1,\ldots,k) \tag{2.119}$$

$$\text{s.t. } \lambda \leq \mu_i \tag{2.120}$$

$$0 \leq \mu_i \leq 1 \qquad (i = 1,\ldots,k) \tag{2.121}$$

$$\lambda, \mathbf{X} \geq 0 \tag{2.122}$$

The above discussion illustrates the main difference between FGP and GP as that of the imprecision associated with the aspiration levels of each objective.

2.7.7 Fuzzy DEA

Sengupta (1992) was the first to explore the use of fuzzy set theory in data envelopment analysis. According to Sengupta (1992) the assumption of deterministic input-output data in DEA limited its scope of application to deterministic systems. Alternatively, the use of stochastic DEA to model non-deterministic systems assumes the data to follow a certain mathematical distribution that is often chosen based on computational convenience rather than on empirical evidence. Sengupta (1992) proposed an approach that combines fuzzy set theory with DEA to incorporate a fuzzy objective function and fuzzy constraints, all of which may be essentially satisfied or dissatisfied. In other words, this approach lends elasticity to the satisfaction of relationships defined by the production plans.

2.7.7.1 The Sengupta Fuzzy DEA Model

Sengupta (1992) fuzzified the primal form of the CCR model with multiple inputs and single output as follows:

$$\text{Find } w = (w_1, w_2, \ldots, w_I) \tag{2.123}$$

$$\text{s.t.} \quad \text{Min } z_c = \Sigma_i\, w_i x_{ic} \tag{2.124}$$

$$\Sigma_i\, w_i x_{ih} \geq y_h \qquad h = (1, 2, \ldots, n) \tag{2.125}$$

$$w_i \geq 0 \qquad i = (1, 2, \ldots, I) \tag{2.126}$$

where n = # of decision-making units; x_{ih} = input i from DMU h; y_h = output from DMU h; and w_i = activity level of input i.

The technical efficiency score of DMU_c is given by:

$$\text{TE}_{\text{CCR}}(\text{X}_{c,}\, \text{Y}_c) = \max\, \text{y}_c/\text{z}_c \tag{2.127}$$

Sengupta (1992) fuzzified this model as:

$$\text{Find } w = (w_1, w_2, \ldots, w_I) \tag{2.128}$$

$$\text{s.t.} \quad \text{Min } \Sigma_i\, w_i x_{ic} \tag{2.129}$$

$$\Sigma_i\, w_i x_{ih}\ \tau\ y_h \qquad h = (1, 2, \ldots, n) \tag{2.130}$$

$$w_i \geq 0 \qquad i = (1, 2, \ldots, I) \tag{2.131}$$

Based on Zimmermann 's (1976) formulation, the fuzzy objective function requires the decision-maker to specify an aspiration level p_c and tolerance limit q_c. Similarly, the fuzzy constraints for each DMU require specification of tolerance limits r_h. The decision-maker must also specify membership functions for the objective function and constraints. For example, the membership functions can be expressed in linear terms as follows.

For the objective function:

$$\mu_c(w) = 1 \qquad \text{if } \Sigma_i w_i x_{ic} \leq p_c \tag{2.132}$$

$$\mu_c(w) = 1 - \Sigma_i w_i x_{ic} - p_c/q_c \qquad \text{if } q_c \leq \Sigma_i w_i x_{ic} \leq p_c + q_c \tag{2.133}$$

$$\mu_c(w) = 0 \qquad \text{if } \Sigma_i w_i x_{ic} \geq p_c + q_c \tag{2.134}$$

For the constraints:

$$\mu_h(w) = 1 \qquad \text{if } \Sigma_i w_i x_{ih} \geq y_h \tag{2.135}$$

$$\mu_h(w) = 1 - y_h - \Sigma_i w_i x_{ih}/r_h \qquad \text{if } y_h \geq \Sigma_i w_i x_{ih} \geq y_h - r_h \tag{2.136}$$

$$\mu_h(w) = 0 \qquad \text{if } y_h - r_h \geq \Sigma_i w_i x_{ih} \tag{2.137}$$

The solution to this fuzzy problem is expressed as: $\lambda_c = \text{Min}_h \{\mu_c(w), \mu_h(w)\}$ and can be found by solving the crisp equivalent which is given as:

$$\text{Max } \lambda_c \tag{2.138}$$

$$\text{s.t.} \quad \lambda_c q_c + \Sigma_i w_i x_{ic} \leq p_c + q_c \tag{2.139}$$

$$\lambda_c r_h - \Sigma_i w_i x_{ih} \leq r_h - y_h \qquad h = (1, 2, \ldots, n) \tag{2.140}$$

$$0 \leq \lambda_c \leq 1 \tag{2.141}$$

$$w_i \geq 0 \qquad i = (1, 2, .., I) \tag{2.142}$$

With the solution to the vector w that maximizes λ the technical efficiency score can be computed as:

$$\text{TE}_{\text{FDEA}} (\text{X}_\text{c}, \text{Y}_\text{c}) = \text{y}_\text{c}/\Sigma_i w_i x_{ic} \tag{2.143}$$

Here, TE represents the fuzzy technical efficiency score for DMU_c.

2.7.7.2 Fuzzy Radial and Non-Radial Models

Girod (1996) introduced fuzziness in the production plans while maintaining crispness in the relationships to be satisfied by the production plans. He used Carlsson's and Korhonen's (1986) framework to formulate the fuzzy BCC and FDH models which are radial measures of efficiency. Only the fuzzy BCC model is reviewed here. The reader is directed to Girod (1996) for a detailed exposition of the two models.

The framework developed by Girod (1996) is based on the assumption that though information regarding inputs and outputs may be imprecise the decision-maker may still be able to specify their upper and lower bounds. Girod's (1996) concept of imprecise production plans differs from Sengupta's (1992) notion of imprecise satisfaction of the relationships to be satisfied by the production plans. Therefore, Girod (1996) allows the inputs to vary between *risk-free* (upper) and *impossible* (lower) bounds and the outputs to vary between *risk-free* (lower) and *impossible* (upper) bounds. Therefore, the fuzzy BCC model is written as:

$$TE_{FBCC}\,(X_c, Y_c) = \text{Min}\ \theta_c \tag{2.144}$$

$$\text{s.t.}\quad [X_c^0 - (X_c^0 - X_c^1)\mu]\theta_c - [X^0 - (X^0 - X^1)\mu]\gamma \geq 0 \tag{2.145}$$

$$[(Y^0 - Y^1)\mu + Y^1)]\gamma \geq (Y_c^1 - Y_c^0)\mu + Y_c^1] \tag{2.146}$$

$$\Sigma_h \gamma_h = 1 \tag{2.147}$$

$$\theta_c,\ \gamma \geq 0 \tag{2.148}$$

Girod (1996) solves this model for various values of $\mu \in [0, 1]$. Mathematically, this translates to the possibility of achieving the production plans corresponding to a particular value of μ.

Parlikar (1996) extended Girod's (1996) work to develop fuzzy non-radial measures of technical efficiency. He formulated fuzzy models for the Färe-Lovell model, the Zieschang model, and the asymmetric Färe model. Only the fuzzy Färe-Lovell model is presented here. The reader is directed to Parlikar (1996) for detailed formulations of the three models.

Maintaining the same assumptions of Girod (1996) the fuzzy Färe-Lovell model formulated by Parlikar (1996) is written as:

$$\text{Min } 1/m\ \Sigma_i\ \theta_i \qquad i = (1, 2, \ldots, I) \tag{2.149}$$

$$\text{s.t.} \quad [X_c^0 - (X_c^0 - X_c^1)\mu]\theta_i - [X^0 - (X^0 - X^1)\mu]\gamma \geq 0 \quad i = (1, 2, \ldots, I), \tag{2.150}$$

$$[(Y^0 - Y^1)\mu + Y^1)]\gamma \geq (Y_c^1 - Y_c^0)\mu + Y_c^1] \tag{2.151}$$

$$\Sigma_h \gamma_h = 1 \tag{2.152}$$

$$\theta_i \leq 1 \qquad i = (1, 2, \ldots, I) \tag{2.153}$$

$$\theta_i, \gamma \geq 0 \qquad i = (1, 2, \ldots, I) \tag{2.154}$$