Chapter 3
Measurement of Banking Efficiency:
Analytical Methods

3.1 Introduction

The main purpose of this chapter is to review the various frontier approaches that have been utilized extensively in the applied research on the banking efficiency. Earlier, the regulators, managers, investors and analysts in the banking sector generally relied on financial accounting ratios to assess the relative efficiency of banks. The main reasons for using ratios as a tool for performance evaluation are to allow comparison among similar sized banks and to control for sector specific characteristics permitting the comparison of individual bank’s ratios with some benchmark for the sector (Halkos and Salamouris, 2004). An inspection of literature provides that many different ratios have been employed to examine various aspects of a bank’s performance. For instance, Intermediation cost, Interest spread, Operating expenditure, Cost to income ratio, Return on Assets, Return on Equity, Business per employee, Income per employee and Business per branch, among others are some commonly used accounting ratios for assessing the financial performance of the banking units (Reserve Bank of India, 2008c).

Though financial accounting ratios are simple to use and relatively easy to understand, but their use to measure bank performance is subject to many criticisms. The financial ratios don’t take account the differences in the business undertaken by different banks, which will in turn be reflected in different combinations of inputs and outputs (Tripe, 2004). DeYoung (1998) suggests that blind pursuit of accounting based benchmarks might reduce a bank’s cost efficiency by cutting back on those expenditures necessary to run the bank properly. Further, Berger et al. (1993a) note that financial ratios may be misleading because they do not control for product mix or input prices. Owing to aforementioned intricacies of the financial accounting ratios, the frontier efficiency analysis gained tremendous popularity in measuring the efficiency of banking industry. Bauer et al. (1998) suggested that frontier efficiency analysis is superior to the financial ratios' analysis. The frontier efficiency approach is based on the recognition that some banks will not be as successful as others in meeting their objectives. The technique measures the performance of each bank in an industry relative to the efficient frontier consisting of dominant banks in the industry. A
bank is classified as fully efficient (with efficiency score equals to 1) if it lies on the frontier and inefficient (with efficiency score ranges from 0 to 1) if its outputs can be produced more efficiently by another set of banks. It is significant to note here that each frontier technique involves various models for deriving a measure of best-practice for the sample of banks and then determine how closely individual banks lie relative to this standard. The best-practice is usually in the form of an efficient frontier that is estimated using econometric or mathematical programming techniques. The frontier techniques summarize bank performance in a single statistic that controls for the differences among banks in a sophisticated multi-dimensional framework that has its roots in economic theory. Further, frontier efficiency measures dominate the traditional ratio analysis in terms of developing meaningful and reliable measures of bank performance. Owing to these features of frontier methodology, the conventional ratio analysis is becoming obsolete.

The available frontier efficiency approaches can be grouped into two major estimation techniques: (i) parametric, and (ii) non-parametric approaches. In parametric approaches, a specific functional form of the production function like Cobb-Douglas and transcendental logarithmic (translog), etc. is required to specify a priori. The efficiency is then assessed in relation to this function with constant parameters and will be different depending on the chosen functional form. On the other hand, non-parametric approaches do not specify a functional form, but nevertheless require certain assumptions about the structure of production technology (e.g., free disposability¹, convexity², etc.). In the non-parametric approaches, a separate mathematical programming problem is needed to solve for obtaining the efficiency scores for the individual banks included in the sample. Further, non-parametric approaches are deterministic in nature since these approaches postulate that all the distances from the efficient frontier are assumed to be caused by inefficiency. Figure 3.1 highlights the major frontier efficiency techniques which are being utilized by the researchers in banking efficiency analyses.
To present the analytical framework for computing the efficiency scores in each frontier approach, the rest of the chapter is organized as follows. Section 3.2 discusses the DEA approach for measuring efficiency in a cross-sectional data setting. Section 3.3 introduces the underlying framework of widely used DEA models in panel data setting. The strengths, limitations, basic requirements and outcomes from a DEA methodology are discussed in Section 3.4. Section 3.5 presents the FDH approach for measuring the efficiency of banks. Section 3.6 discusses the parametric efficiency measurement approach of SFA. Further, the details on the DFA, TFA and RTFA approaches are given in the Section 3.7. In Section 3.8, a comparative analysis of DEA and SFA techniques is presented. In particular, this section focuses on why both the approaches produce different estimates of bank’s efficiency. The final section concludes the discussion.

3.2 Data envelopment analysis (DEA)

DEA is a linear (mathematical) programming based method first originated in the literature by Charnes et al. (1978) as a reformulation of the Farrell’s (1957) single-output, single-input radial measure of technical efficiency to the multiple-output, multiple-input case. The subsequent developments in DEA are very extensive. Interested parties are directed to those provided by Seiford and Thrall (1990), Ali and Seiford (1993), Charnes et al. (1994), Seiford (1996), Zhu (2003),
Ray (2004) and Copper et al. (2007). What follows is a general discussion of DEA with primary attention directed to describe a few widely used DEA models. DEA calibrates the level of technical efficiency (TE) on the basis of an estimated discrete piecewise frontier (or so called efficient frontier or best-practice frontier or envelopment surface) made up by a set of Pareto-efficient decision making units (DMUs)\(^3\). In all instances, these Pareto-efficient banks located on the efficient frontier, compared to the others, use minimum productive resources given the outputs (input-conserving orientation), or maximize the output given the inputs size (output-augmenting orientation), and are called the best-practice performers or reference units or peer units within the sample of banks. These Pareto-efficient banks have a benchmark efficiency score of unity that no individual bank’s score can surpass. In addition, it is not possible for the Pareto-efficient unit to improve any input or output without worsening some other input or output. It is significant to note that the efficient frontier provides a yardstick against which to measure the relative efficiency of all other banks that do not lie on the frontier. The banks which do not lie on the efficient frontier are deemed relatively inefficient (i.e., Pareto non-optimal banks) and receives a TE score between 0 and 1. The efficiency score of each bank can be interpreted as the radial distance to the efficient frontier. In short, the DEA forms a non-parametric surface frontier (more formally a piecewise-linear convex isoquant) over the data points to determine the efficiency of each bank relative to this frontier.

Using actual data for the banks under consideration, DEA employs linear programming technique to construct efficient or best-practice frontier. In fact, a large number of linear programming DEA models have been proposed in the literature to compute efficiency of individual banks corresponding to different technical or behavioural goals. Essentially, each of these various models seek to establish which of \( n \) banks determine the envelopment surface or best-practice frontier or efficient frontier. The geometry of this surface is prescribed by the specific DEA model employed. Nevertheless, for analytical purpose, we can classify DEA models used in banking efficiency models in two broad categories: (i) Non-allocation DEA models, and (ii) Allocation DEA models.

**3.2.1 Non-allocation DEA models**

The non-allocation DEA models compute the relative TE scores for individual banks without using any information on prices of inputs and outputs.
Before discussing the methods for efficiency measurement it is necessary to look at the different perspectives of technical efficiency. Technical efficiency (TE) refers to the conversion of physical inputs, such as labour and capital, into outputs relative to best-practice. TE, thus, relates to the productivity of inputs (Sathye, 2001). It is a comparative measure of how well it actually processes inputs to achieve its outputs, as compared to its maximum potential for doing so, as represented by its production possibility frontier (Barros and Mascarenhas, 2005). Accordingly, TE of the bank is its ability to transform multiple resources into multiple financial services (Bhattacharyya et al., 1997b). A bank is said to be technically inefficient if it operates below the frontier. A measure of TE helps to determine inefficiency due to the input/output configuration as well as the size of operations.

Charnes et al. (1994) described three possible orientations in DEA models for computing TE scores: (i) Input-oriented models are the models where banks are deemed to produce a given amount of outputs with the minimum possible amount of inputs (inputs are controllable). In this orientation, the inefficient banks are projected onto the efficient frontier by decreasing their consumption of inputs. Input minimization allows us to determine the extent to which a bank can reduce inputs while maintaining the current level of outputs; (ii) Output-oriented models are models where banks are deemed to produce with given amounts of inputs the maximum possible amount of outputs (outputs are controllable). In this orientation, inefficient banks are projected onto the efficient frontier by increasing their production of outputs. Output maximization might be used when the inputs are constrained, and emphasis is on increasing the outputs; and (iii) Base-oriented models (or additive or non-oriented models) are models where banks are deemed to produce the optimal mix of inputs and outputs (both inputs and outputs are controllable). Here, the inefficient banks are projected onto the efficient frontier by simultaneously reducing their inputs and increasing their outputs to reach an optimum level.

Figure 3.2 describes the different orientations used in DEA framework using the simple case of a single-input and single-output production system. QQ' represents efficient frontier, and Bank D is an inefficient unit. Point I constitutes the benchmark for inefficient Bank D in the input-oriented model. The relative
efficiency of Bank D is given by the ratio of distances $D_i/D_1$. Point O is the projection of D in the output-oriented model. The relative efficiency of Bank D is then $DD_o/OD_o$. Finally, point B is the base-projection of Bank D in the base-oriented model.

![Figure 3.2: Orientations in DEA](image)

In the empirical studies, the researchers have widely utilized the input-oriented and output-oriented models. An illustration of the TE measurement from input-oriented perspective is provided in Figure 3.3. The figure illustrates a two-dimension efficient frontier in input space (i.e., an isoquant $L(y)$) in which all four banks (A, B, C and D) produce the same amount of output $y$, but with varying amounts of inputs $x_1$ and $x_2$. The efficient frontier in input space is defined by banks A, B, C and D that require minimum inputs to produce the same level of output. These units are labeled as efficient banks and have TE score equal to 1. On the other hand, banks E and F are inefficient because both require more of each input to produce the same amount of output. In input-oriented context, a measure of TE for an inefficient bank can be defined as:

$$\theta_{\text{input}} = \frac{\text{Minimum input}}{\text{Actual input}}$$

The measure of TE for Bank E is defined as $\theta^E = OE/OE$, respectively. It is significant to note that $\theta^U$ is less than 1. Further, the inefficient Bank E can
move on to the efficient frontier (and in a way get the status of efficient bank in Farrell’s sense) by a radial (or proportional) reduction in inputs by amount EE'.

Figure 3.3: Input-oriented technical efficiency

Figure 3.4 depicts the output-oriented measure of technical efficiency. In this case, the banks A, B, C, D, E and F produce any combination of the two outputs $y_1$ and $y_2$ that fall within the production set $P(x)$ using a given amount of inputs. The piecewise linear boundary ABCD is the locus of efficient production and, therefore, banks A, B, C and D are rated as efficient. Banks E and F lies within the production possibility set and are, therefore, rated inefficient. In output-oriented context, TE is defined as the proportion to which outputs can be expanded radially without changing the input level. A measure of TE for an inefficient bank can be defined as:

$$\theta_{output} = \frac{\text{Actual output}}{\text{Maximum output}}$$

To derive the efficiency of Bank E, we simply calculate how far E can be moved towards the frontier along the dotted line through the origin. The measures of TE for Bank E is defined as $\theta^E_{output} = OE/OB$, respectively. Further, the inefficient Bank E can move on to the efficient frontier (and in a way get the status of efficient bank in Farrell’s sense) by a radial (or proportional) augmentation in outputs by amount EB.
The widely used non-allocation DEA models to compute technical efficiency scores are the CCR model, the BCC model, the Additive model, the Multiplicative model and the Slack-based measures (SBM) model. Besides this, the researchers used extensions of CCR and BCC models for specific purposes like ranking of banks, incorporating value judgments, including non-discretionary inputs and outputs, etc. The following sub-sections discuss various non-allocation DEA models.

3.2.1.1 The CCR model

In their seminal paper entitled, “Measuring the efficiency of decision making units”, which is published in European Journal of Operational Research, Charnes et al. (1978) developed a DEA model which got tremendous popularity with the name CCR DEA model. The CCR model is based on the assumptions of constant returns-to-scale (CRS), strong disposability of inputs and outputs, and convexity of the production possibility set. The application of CCR model not only provides technical efficiency scores for individual banks but also provides vital information on input and output slacks, and reference set for inefficient banks. The literature spells out that there are two distinct variants of the CCR model: input-oriented CCR model (CCR-I), and output-oriented CCR model (CCR-O).

3.2.1.1.1 CCR-I

To illustrate input-oriented CCR DEA model, consider a set of $n$ banks ($j = 1,\ldots,n$), utilizing quantities of inputs $x \in \mathbb{R}_+^m$ to produce quantities of
outputs \( y \in R^s_+ \). We can denote \( x_{ij} \) the amount of the \( i \)th input used by the bank \( j \) \((i = 1, \ldots, m)\) and \( y_{rj} \) the amount of the \( r \)th output produced by the bank \( j \) \((r = 1, \ldots, s)\). In the CCR model, the multiple-inputs and multiple-outputs of each bank are aggregated into a single virtual input and virtual output, respectively. The input-oriented TE score for target bank ‘o’ can be obtained by solving the following fractional programming model:

\[
\max_{u,v} \ h_o(u,v) = \frac{\text{Virtual Output}_o}{\text{Virtual Input}_o} = \frac{\sum_{r=1}^s u_r y_{ro}}{\sum_{i=1}^m v_i x_{io}}
\]

subject to

\[
\sum_{r=1}^s u_r y_{ro} \leq 1 \\
\sum_{i=1}^m v_i x_{io} \\
u_r \geq \epsilon \\
v_i \geq \epsilon
\]

where \( y_{ro} \) = the amount of the \( r \)th output produced by the bank ‘o’, \( x_{io} \) = the amount of the \( i \)th input used by the bank ‘o’, \( u_r \) = the weight given to output \( r \), \( v_i \) = the weight given to input \( i \), \( \epsilon \) = a non-Archimedean (infinitesimal) constant.

The objective of this model is to determine positive and unknown input and output weights that maximize the ratio of a virtual output to a virtual input for bank ‘o’. The constraints restrict that the ratio of virtual output to the virtual input for each bank to be less than or equal to 1. This implies that the maximal efficiency, \( h_o^* \), is at the most equal to 1. The justification for \( \epsilon \) is twofold: first, to ensure that the denominator is never zero, and second, to ensure that each input and output is considered. It is important to note that the optimal output and input weights (i.e., \( u_r^* \) and \( v_i^* \)) are obtained through optimization (i.e., linear programming solution). Such optimization is performed separately for each bank in order to compute the weights and efficiency scores.

Charnes and Cooper (1962) developed a transformation from a fractional programming problem to an equivalent linear programming problem. By using the transformation of the variables,
\[
\mu_r = tu_r, \\
\nu_i = tv_i \\
t = \frac{1}{\sum_{i=1}^{m} V_i x_{io}}
\]

The fractional CCR model (3.1) can be transformed into the following linear programming model, which is popularly known as ‘multiplier form’ of CCR model:

\[
\begin{align*}
\text{max } & \quad f_0(\mu) = \sum_{r=1}^{s} \mu_r y_{ro} \\
\text{subject to } & \quad \sum_{i=1}^{m} V_i x_{io} = 1 \\
& \quad \sum_{r=1}^{s} \mu_r y_{oj} - \sum_{i=1}^{m} V_i x_{io} \leq 0, \quad j = 1,...,n \\
& \quad \mu_r \geq \varepsilon, \quad r = 1,...,s \\
& \quad \nu_i \geq \varepsilon, \quad i = 1,...,m
\end{align*}
\]

The dual program of the model (3.2), which is popularly known as ‘envelopment form’ of CCR model is given as:

\[
\begin{align*}
\text{min } & \quad g_0(\theta^{CCR}, s^+, s^-) = \theta^{CCR} - \varepsilon \left( \sum_{i=1}^{n} s_i^+ + \sum_{j=1}^{m} s_j^- \right) \\
\text{subject to } & \quad \sum_{j=1}^{n} \lambda_j x_{ij} + s_i^- = \theta^{CCR} x_{io}, \quad i = 1,...,m \\
& \quad \sum_{j=1}^{n} \lambda_j y_{oj} - s_j^+ = y_{ro}, \quad r = 1,...,s \\
& \quad \lambda_j \geq 0, \quad j = 1,...,n \\
& \quad s^+, s^- \geq 0 \\
& \quad 0 < \varepsilon \leq 1
\end{align*}
\]

The primal model has \(n + s + m + 1\) constraints while the dual has \(m + s\) constraints. The number of banks \((n)\) should usually be considered larger than the number of inputs and outputs \((m + s)\) in order to provide a fair degree of discrimination of results. In view of this, it is clear that dual model (3.3) will be simpler to solve as it has \(n + 1\) fewer constraints than the primal model (3.2). It should be noted that both the primal (multiplier form) and dual (envelopment form) problems have the same solutions.
The scalar $\theta^{CCR}$, corresponding to bank $o$’s TE score, represents the largest possible radial contraction that is proportionally applied to the bank $o$’s inputs in order to project it to a point on the efficient frontier that corresponds to the minimal consumption of inputs $\left( \sum_{j=1}^{n} \lambda_{j} x_{ij} + s_{i}^{-} \right)$ required to produce bank $o$’s current output levels $\left( \sum_{j=1}^{n} \lambda_{j} y_{ij} - s_{r}^{+} \right)$. For inefficient banks the value of $\theta^{CCR} < 1$ represents the proportion of inputs that the bank should be using to produce its current levels of outputs, such that $1 - \theta^{CCR}$ corresponds to bank $o$’s level of inefficiency. The value of $\theta^{CCR}$ is limited to be $0 < \theta^{CCR} \leq 1$. Any non-zero values of $\lambda_{j}$ indicate that an efficient bank is in the reference set of bank ‘$o$’. The $s_{i}^{-}$ slack term equals the input access that remains in input $i$ of bank ‘$o$’ after the radial contraction was applied bank $o$’s inputs, and the $s_{r}^{+}$ slack term equals the shortfall in the production of output $r$. The $\varepsilon$ term in the objective function represents a small positive number ($10^{-6}$) whose purpose is to maximize the sum of the slacks should more than one optimal solution exists. Doing so, however, can lead to some theoretical difficulties that can be avoided by solving the CCR model in two stages. In Stage 1, the model is solved for the optimal value of $\theta^{CCR}$ (i.e., $\theta^{CCR}$), while in Stage 2, the value of $\theta^{CCR}$ is fixed to $\theta^{CCR}$ and the model is solved such that it maximizes the values of the slacks. Thus, the computation proceeds in two stages: with maximal reduction of inputs achieved first, via the optimal $\theta^{CCR}$; then in the second stage, substituting the $\theta^{CCR}$ in the dual constraints, movement onto the efficient frontier is achieved via the slack variables $(s_{i}^{-}$ and $s_{r}^{+})$. These stages are outlined below.

Stage 1 focuses on obtaining the TE scores in Farrell-Debreu’s sense by ignoring the presence of non-zero slacks. For getting TE score for bank ‘$o$’, the model (3.4) is to be solved:
\[ \theta^{CCR}_o = \min \theta^{CCR} \]

subject to
\[ \sum_{j=1}^{n} \lambda_j x_{ij} \leq \theta x_{io} \quad i = 1, \ldots, m \quad (3.4) \]
\[ \sum_{j=1}^{n} \lambda_j y_{ij} \geq y_{io} \quad r = 1, \ldots, s \]
\[ \lambda_j \geq 0, \quad j = 1, \ldots, n \]

\( \theta^{CCR}_o \) represent the input-oriented TE score of bank ‘o’. After calculating model (3.4), we obtain input and output slack values as
\[ s_i^+ = \theta^{CCR}_o x_{io} - \sum_{j=1}^{n} \lambda_j x_{ij} \quad i = 1, \ldots, m \]
\[ s_r^+ = \sum_{j=1}^{n} \lambda_j y_{ij} - y_{ro} \quad r = 1, \ldots, s \]

where \( s_i^- \) and \( s_r^+ \) represent input and output slacks, respectively.

In Stage 2, we optimize the slacks by fixing \( \theta^{CCR}_o \) in the following linear programming problem:
\[ \max \left( \sum_{r=1}^{s} s_r^+ + \sum_{i=1}^{m} s_i^- \right) \]

subject to
\[ \sum_{j=1}^{n} \lambda_j x_{ij} + s_i^- = \theta^{CCR}_o x_{io}, \quad i = 1, \ldots, m \quad (3.5) \]
\[ \sum_{j=1}^{n} \lambda_j y_{ij} - s_r^+ = y_{ro}, \quad r = 1, \ldots, s \]
\[ \lambda_j \geq 0, \quad j = 1, \ldots, n \]

The solution of the model (3.5) yields optimal values of input and output slacks \( s_i^- \) and \( s_r^+ \).

The interpretation of the results of envelopment model (3.3) can be summarized as:

(1) The bank ‘o’ is Pareto-efficient if and only if \( \theta^{CCR}_o = 1 \) and \( s_i^- = s_r^+ = 0 \) for all \( i \) and \( r \). Otherwise, if \( \theta^{CCR}_o < 1 \) then the bank ‘o’ is inefficient, i.e., the bank ‘o’ can either increase its output levels or decrease its input levels.

(2) The left-hand side of the envelopment model is usually called the ‘reference set’, and the right-hand side represents a specific bank under evaluation. The non-
zero optimal $\lambda_j^*$ represents the benchmarks for a specific bank under evaluation. The reference set provides coefficients ($\lambda_j^*$) to define the hypothetical efficient bank. The reference set or the efficient target shows how inputs can be decreased and outputs increased to make the bank under evaluation efficient.

### 3.2.1.1.2 CCR-O

The output-oriented CCR model focuses on maximal movement via proportional augmentation of output for a given level of inputs. To drive CCR-O model, we minimize the inefficiency of bank ‘o’ given by the ratio of virtual input to virtual output under the constraints that so defined inefficiency cannot be lower than one for itself or for any of the other banks. The required optimization problem is:

$$\min_{u,v} z_o(u,v) = \frac{\text{Virtual Input}_{o}}{\text{Virtual Output}_{o}} = \frac{\sum_{i=1}^{m} v_i x_{io}}{\sum_{r=1}^{s} u_r y_{ro}}$$

subject to

$$\sum_{i=1}^{m} v_i x_{ij} \geq 1 \quad j = 1, \ldots, n \quad (3.6)$$

$$\sum_{r=1}^{s} u_r y_{rj} \quad r = 1, \ldots, s$$

$$u_r \geq \varepsilon \quad r = 1, \ldots, s$$

$$v_i \geq \varepsilon \quad i = 1, \ldots, m$$

Again the Charnes-Cooper (1962) transformation for fractional programming yields the linear programming model, popularly known as output-oriented multiplier (primal) and envelopment (dual) models, as follows:

<table>
<thead>
<tr>
<th>Model orientation</th>
<th>Multiplier model (Primal)</th>
<th>Envelopment model (Dual)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Output-oriented</td>
<td>$\min_{\mu,\nu} w_o(v) = \sum_{i=1}^{m} v_i x_{io}$ subject to $\sum_{r=1}^{s} \mu_r y_{ro} = 1$ $\sum_{i=1}^{m} v_i x_{ij} - \sum_{r=1}^{s} \mu_r y_{rj} \geq 0$ $\mu_r, v_i \geq \varepsilon$</td>
<td>$\max_{\phi, \check{\lambda}^+, \check{s}^+} h_o(\phi^{CCR}, \check{s}^+, \check{s}^-) = \phi^{CCR} + \varepsilon \left( \sum_{i=1}^{m} x_{io}^{+} + \sum_{j=1}^{n} \check{s}<em>j^- \right)$ subject to $\sum</em>{j=1}^{n} \check{\lambda}<em>j x</em>{ij} + \check{s}<em>j^- = x</em>{io}$ $\sum_{j=1}^{n} \check{\lambda}<em>j y</em>{rj} - \check{s}<em>r^+ = \phi^{CCR} y</em>{ro}$ $\check{\lambda}_j, \check{s}_j^-, \check{s}_r^+ \geq 0$</td>
</tr>
</tbody>
</table>
Like model (3.3), the output-oriented envelopment CCR model is also solved in a two-stage process. First, we calculate $\phi^{CCR}$ by ignoring the slacks. Then we optimize the slacks by fixing $\phi^{CCR}$ in the following linear programming problem:

$$
\text{max} \left( \sum_{r=1}^{s} s_r^+ + \sum_{i=1}^{m} s_i^- \right) \\
\text{subject to} \\
\sum_{j=1}^{n} \lambda_j x_{ij} + s_i^- = x_{i0} \quad i = 1, \ldots, m \quad (3.7) \\
\sum_{j=1}^{n} \lambda_j y_{rj} - s_r^+ = \phi^{CCR} y_{ro} \quad r = 1, \ldots, s \\
\lambda_j \geq 0, \quad j = 1, \ldots, n
$$

### 3.2.1.2 The BCC model

The BCC model has been developed by Banker et al. (1984) as an extension of the CCR model to allow for returns-to-scale to be variable. Thus, BCC model computes efficiency scores corresponding to the assumption of variable returns-to-scale (VRS). It is a more flexible than the CCR model since it allows for constant, increasing, and decreasing returns-to-scale. Banker et al. (1984) showed that solutions to CCR and BCC models allow a decomposition of technical efficiency (TE) into pure technical efficiency (PTE) and scale efficiency (SE) components. Like the CCR model, BCC model also has two variants: input-oriented BCC model (BCC-I) and output-oriented BCC model (BCC-O).

#### 3.2.1.2.1 BCC-I

BCC-I model measures the pure technical efficiency of the bank ‘o’ by solving the following pair of primal (multiplier form) and dual (envelopment) linear programming models:
Table 3.2: Input-oriented BCC model (BCC-I)

<table>
<thead>
<tr>
<th>Model orientation</th>
<th>Multiplier model (Primal)</th>
<th>Envelopment model (Dual)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Input-oriented</td>
<td>max $f_o(\mu, \mu_o) = \sum_{r=1}^n \mu_r y_{or} - \mu_o$ subject to $\sum_{r=1}^n v_r x_{or} = 1$ $\sum_{r=1}^n \mu_r y_r - \sum_{r=1}^n v_r x_r - \mu_o \leq 0$ $\mu_r, v_r \geq 0$ $\mu_o$ free in sign</td>
<td>$\min \theta_{BCC} = 1 - \sum_{r=1}^n \sum_{s=1}^n \sum_{t=1}^n (s_r^t - \theta_{BCC} x_{or})$ subject to $\hat{\lambda}<em>j x</em>{or} + s_r^t = \theta_{BCC} x_{or}$ $\sum_{r=1}^n \hat{\lambda}<em>j y_r - s_r^t = y_r$ $\sum</em>{j=1}^n \hat{\lambda}_j = 1$ $\hat{\lambda}_j, s_r^t, s_r^t \geq 0$</td>
</tr>
</tbody>
</table>

Above models differ from their CCR counterparts in the free variable, $\mu_o$, in the primal model and the constraint $\sum_{j=1}^n \hat{\lambda}_j = 1$, in the dual model. It is worth noting that the convexity constraint, $\sum_{j=1}^n \hat{\lambda}_j = 1$, essentially ensures that an inefficient bank is only ‘benchmarked’ against banks of a similar size. The free variable, $\mu_o$, relaxes the constant returns-to-scale condition by not restricting the envelopment surface to go through the origin. The BCC model can be solved using a two-phased approach similar to that for the CCR model. The first phase provides $\theta_{o,bcc}^*$ and then $\theta_{o,bcc}^*$ is used in the second phase to solve for the input excesses, $s_{r}^*$, and output shortfalls, $s_{r}^*$. Because the BCC model imposes an additional constraint, $\sum_{j=1}^n \hat{\lambda}_j = 1$, the feasible region of the BCC model is a subset of that of the CCR model. The relationship between the optimal objective values of the CCR and BCC models is that $\theta_{o,bcc}^* \geq \theta_{o,CCR}^*$. Therefore, a bank found to be efficient with the CCR model will also be found to be efficient with the corresponding BCC model. A measure of SE for bank ‘o’ can be obtained as a ratio of efficiency measure from CCR-I model to efficiency measure from BCC-I model, i.e., $\theta_{o,CCR}^*/\theta_{o,bcc}^*$.  

81
3.2.1.2.2 BCC-O

The output-oriented BCC model measures the efficiency of the bank ‘o’ by solving the following pair of primal (multiplier form) and dual (envelopment) linear programming models:

Table 3.3: Output-oriented BCC model (BCC-O)

<table>
<thead>
<tr>
<th>Model orientation</th>
<th>Multiplier model (Primal)</th>
<th>Envelopment model (Dual)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Output-oriented</td>
<td>[ \begin{align*} \min \ w_o(\nu, \nu_o) &amp;= \sum_{i=1}^{m} v_i x_{io} + v_o \ \text{subject to} &amp; \sum_{i=1}^{s} \mu_i y_{io} = 1 \ &amp; \sum_{i=1}^{s} v_i x_{io} - \sum_{j=1}^{n} \mu_j y_{ij} + v_o \geq 0 \ &amp; \mu, v_i \geq \varepsilon \ &amp; v_o \text{ free in sign} \end{align*} ]</td>
<td>[ \begin{align*} \max \phi^{BCC} \quad &amp; + \epsilon \left( \sum_{j=1}^{s} s_i^+ + \sum_{i=1}^{m} s_j^- \right) \ \text{subject to} &amp; \sum_{j=1}^{n} \lambda_j x_{ij} + s_i^- = x_{io} \ &amp; \sum_{j=1}^{n} \lambda_j y_{ij} - s_i^+ = \phi^{BCC} y_{ro} \ &amp; \sum_{j=1}^{n} \lambda_j = 1 \ &amp; \lambda_j, s_i^+, s_i^- \geq 0 \end{align*} ]</td>
</tr>
</tbody>
</table>

Again, BCC-O models illustrated in the Table 3.3 differ from their CCR counterparts in the free variable, \( \nu_o \), in the primal model, and the constraint, \( \sum_{j=1}^{n} \lambda_j = 1 \), in the dual model. The free variable, \( \nu_o \) relaxes the assumption of constant returns-to-scale by not restricting the envelopment surface to go through the origin. The BCC-O model can also be solved using a two-phased approach similar to that for the CCR-O model. The first phase provides \( \phi^{BCC}_o \) and then \( \phi^{BCC}_o \) is used in the second phase to solve for the input excesses, \( s_i^- \), and output shortfalls, \( s_i^+ \). A measure of SE for bank ‘o’ can be obtained as a ratio of efficiency measure from CCR-O model to efficiency measure from BCC-O model, i.e., \( \phi^{CCR}_o / \phi^{BCC}_o \).

3.2.1.3 Additive model

In the preceding models (CCR and BCC), the projection of inefficient banks to the envelopment surface is based on the model orientation. As CCR-I (or BCC-I) model focuses on radial movement toward the frontier through the proportional reduction of inputs, while CCR-O (or BCC-O) models does this through proportional augmentation of outputs. Charnes et al. (1985) introduced the Additive
or Pareto–Koopmans (PK) model which provides a non-oriented measure that simultaneously reduces the inputs and augments the outputs by taking the slacks into account when measuring efficiency. The envelopment surface in the additive model is similar to BCC model in that it allows for scale effects creating the same VRS efficiency frontier. This is due to presence of the convexity constraint \( \sum_{j=1}^{n} \lambda_j = 1 \) in the dual (envelopment form), and equivalently, \( \mu_o \) in the primal (multiplier form) problem.

The primal (multiplier) and dual (envelopment) form of linear programming problems of the additive model can be expressed as follows:

<table>
<thead>
<tr>
<th>Model orientation</th>
<th>Multiplier model (Primal)</th>
<th>Envelopment model (Dual)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Non-oriented</td>
<td>( \min_{\mu, \nu} \quad w_o = \sum_{i=1}^{s} \mu_i y_{ro} - \sum_{j=1}^{m} \nu_j x_{io} + \mu_o )</td>
<td>( \max_{\lambda, \nu} \quad g_o (\lambda, s_i^+, s_r^+) = -\sum_{r=1}^{s} s_r^+ - \sum_{j=1}^{m} s_j^- )</td>
</tr>
<tr>
<td></td>
<td>subject to</td>
<td>subject to</td>
</tr>
<tr>
<td></td>
<td>( \sum_{r=1}^{s} \mu_i y_{ij} - \sum_{j=1}^{m} \nu_j x_{ij} - \mu_o \leq 0 )</td>
<td>( \sum_{j=1}^{n} \lambda_j y_{ij} - s_r^+ = y_{ro} )</td>
</tr>
<tr>
<td></td>
<td>( \mu_r \geq 1 )</td>
<td>( \sum_{j=1}^{n} \lambda_j x_{ij} + s_j^- = x_{io} )</td>
</tr>
<tr>
<td></td>
<td>( \nu_j \geq 1 )</td>
<td>( \sum_{j=1}^{n} \lambda_j = 1 )</td>
</tr>
<tr>
<td></td>
<td>( \mu_o ) free in sign</td>
<td>( \lambda, s_i^-, s_r^+ \geq 0 )</td>
</tr>
</tbody>
</table>

In the additive model, the bank ‘o’ is efficient if and only if \( w_o = g_o \). In other words, bank ‘o’ is efficient if \( s_i^0 = 0 \) and \( s_r^1 = 0 \). If any component of the slack variables is positive then it is inefficient, and the values of non-zero components identify the sources and amounts of inefficiency in the corresponding inputs and outputs. The solution to envelopment model gives the optimal values \( s_i^{\ast} \) and \( s_r^{\ast} \), which can be used to define

\[
\hat{y}_{ro} = y_{ro} + s_r^{\ast} \geq y_{ro} \\
\hat{x}_{io} = x_{io} - s_i^{\ast} \leq x_{io}
\]

Note that the slacks are all independent of each other. Hence, an optimum is not reached until it is not possible to increase an output \( \hat{y}_{ro} \) or reduce an input \( \hat{x}_{io} \) without decreasing some other output or increasing some other input.
3.2.1.4 Multiplicative model

In the preceding DEA models, efficiency is viewed as the sum of outputs divided by the sum of inputs. This means that adding one more output results in added input without any effect on the other outputs. However, in some processes output levels (or input levels) may be interdependent (Sherman, 1988). Charnes et al. (1982) introduced an alternative formulation of DEA known as ‘multiplicative model’ which provides a measure of efficiency based on the ratio of the weighted multiplicative product of outputs divided by the weighted multiplicative product of inputs in order to account for interdependencies between input or output levels. The input \( (v_i) \) and output \( (u_r) \) weights are applied as powers to the input and output variables as can be seen in the multiplicative formulation below:

\[
\max \quad \prod_{r=1}^{s} \frac{y_{ro}^{u_r}}{\prod_{i=1}^{m} x_{io}^{v_j}}
\]

subject to

\[
\prod_{r=1}^{s} \frac{y_{ro}^{u_r}}{\prod_{i=1}^{m} x_{io}^{v_j}} \leq 1, \quad j = 1, \ldots, n \tag{3.8}
\]
\[
u_r \geq 1, \quad r = 1, \ldots, s
\]
\[
v_i \geq 1, \quad i = 1, \ldots, m
\]

Model (3.8) can be converted to a linear programming problem by taking logarithms, and its primal (multiplier) and envelopment (dual) formulations are given as:

<table>
<thead>
<tr>
<th>Table 3.5: Multiplicative model</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Multiplier model</strong> (Primal)</td>
</tr>
<tr>
<td>max ( \sum_{r=1}^{s} u_r \ln(y_{ro}) - \sum_{i=1}^{m} v_i \ln(x_{io}) )</td>
</tr>
<tr>
<td>subject to ( \sum_{r=1}^{s} u_r \ln(y_{ro}) - \sum_{i=1}^{m} v_i \ln(x_{io}) \leq 0 )</td>
</tr>
<tr>
<td>( u_r \geq 1 )</td>
</tr>
<tr>
<td>( v_i \geq 1 )</td>
</tr>
<tr>
<td><strong>Envelopment model</strong> (Dual)</td>
</tr>
</tbody>
</table>

84
Thus, the envelopment surface in the Multiplicative model is piecewise log-linear instead of piecewise linear, which is the envelopment surface for the other DEA models. As with the Additive model, a bank ‘o’ is only considered to be efficient if all its slacks are zero (Cooper et al., 2007). The above model is also called Variant Multiplicative model, which has a constant returns-to-scale envelopment surface. The Invariant Multiplicative model has the same formulation for the primal and the dual except that the convexity constraint in the dual and the variable $\mu_o$ in the primal are added to the model. As a result, the envelopment surface will be variable returns-to-scale.

### 3.2.1.5 Non-radial slack-based measures (SBM) model

The standard CCR and BCC DEA models, so defined, are based on the proportional reduction (augmentation) of input (output) vectors and do not take account of slacks. While the additive DEA model can capture slacks but it is neither units-invariant nor able to generate a scalar measure of efficiency. Tone (2001) introduced the non-radial or slacks-based measure (SBM) model to deal with inputs/outputs individually, contrary to the radial approaches that assume proportional changes in inputs/outputs. This scalar measure deals directly with the input excesses and the output shortfalls of the concerned banks. It is invariant to the units of measurement and is monotone increasing in each input and output slack. Furthermore, it is reference-set dependent, i.e., the measure is determined only by its reference set and is not affected by statistics over the whole data set.

As far as the orientations in SBM model are concerned, we have input-oriented, output-oriented and non-oriented models. The linear programming problems of the SBM models corresponding to three model orientations are given below:
Table 3.6: SBM models

<table>
<thead>
<tr>
<th>Model orientation</th>
<th>Input-oriented</th>
<th>Output-oriented</th>
<th>Non-oriented</th>
</tr>
</thead>
<tbody>
<tr>
<td>min $\theta = 1 - \frac{1}{m} \sum_{j=1}^{m} s_{i}^{-} / x_{i0}$</td>
<td>min $\phi = \frac{1}{1 + \frac{1}{s} \sum_{r=1}^{s} s_{r}^{+} / y_{ro}}$</td>
<td>min $\rho = \frac{1}{1 + \frac{1}{s} \sum_{r=1}^{s} s_{r}^{+} / y_{ro}}$</td>
<td></td>
</tr>
<tr>
<td>subject to $\sum_{j=1}^{n} \lambda_{j} x_{ij} + s_{i}^{-} = x_{i0}$</td>
<td>subject to $\sum_{j=1}^{n} \lambda_{j} y_{ij} - s_{r}^{+} = y_{ro}$</td>
<td>subject to $\sum_{j=1}^{n} \lambda_{j} x_{ij} + s_{i}^{-} = x_{i0}$</td>
<td></td>
</tr>
<tr>
<td>$\sum_{j=1}^{n} \lambda_{j} y_{ij} - s_{r}^{+} = y_{ro}$</td>
<td>$\sum_{j=1}^{n} \lambda_{j} y_{ij} - s_{r}^{+} = y_{ro}$</td>
<td>$\sum_{j=1}^{n} \lambda_{j} y_{ij} - s_{r}^{+} = y_{ro}$</td>
<td></td>
</tr>
<tr>
<td>$\lambda_{j}, s_{i}^{-}, s_{r}^{+} \geq 0$</td>
<td>$\lambda_{j}, s_{i}^{-}, s_{r}^{+} \geq 0$</td>
<td>$\lambda_{j}, s_{i}^{-}, s_{r}^{+} \geq 0$</td>
<td></td>
</tr>
</tbody>
</table>

Note that both the traditional radial and non-radial SBM DEA models yield the same frontier, but may yield different efficient targets even when the envelopment models do not have non-zero slacks. The objective function in the SBM model satisfies unit invariant because the numerator and denominator are measured in the same units for each bank in the above equation. Furthermore, the non-radial efficiency score also lies between 0 and 1.

### 3.2.2 Extensions of basic non-allocation DEA models

#### 3.2.2.1 Super-efficiency models

When a DMU under evaluation is not included in the reference set of the envelopment models, the resulting DEA models are called super-efficiency DEA models (Zhu, 2003). The first super-efficiency model has been developed by Andersen and Petersen (1993) to provide strict ranking to all DMUs in the sample. Their idea is explained in the Figure 3.5. Figure illustrates that A, B, C and D are efficient banks which make up the best-practice frontier. The inefficient Bank E is compared to a reference point which is the linear combination of the nearest peers on the efficient frontier. In case of Bank E, the reference (or virtual) point, E’, is the linear combination of C and D. The efficiency score of E is OE/OE’ which less than one. The efficiency score of an inefficient bank remains same under super-efficiency and standard DEA approach. The difference exists only when it comes to an efficient bank.
Now consider an efficient bank, say Bank B. The efficiency score of Bank B under the standard DEA approach is \( OB/OB = 1 \), while the efficiency score under the super-efficiency model is determined by excluding B from the original reference set (line ABCD) and then compare B to the new reference set (line ACD) formed by the remaining efficient banks. Thus, the efficiency score of Bank B under the super-efficiency model will be \( OB'/OB \), which is greater than 1. This implies that even proportional increase in input, B can still remain as an efficient bank. Thus, in the super-efficiency model, all the relative efficient banks would have an efficiency score equal to or greater than 1. This procedure makes the ranking of efficient banks possible (i.e., higher super-efficiency score implies higher rank). However, the inefficient units which are not on the efficient frontier, and with an initial DEA score of less than 1, would find their relative efficiency score unaffected by their exclusion from the reference set of banks.

Later, Thrall (1996), Dula’ and Hickman (1997), Seiford and Zhu (1999), Xue and Harker (2002), Tone (2002a), Lovell and Rouse (2003), and Bogetoft and Hougaard (2004) show the infeasibility problems in the Andersen and Petersen (A-P) super-efficiency model. To deal with infeasible problems, Tone (2002a) provides a super-efficiency model using the slacks-based measure of efficiency, which is non-radial and deals with input/output slacks directly.

The linear programming problems for the super-efficiency models as developed by Andersen and Petersen (1993) and Tone (2002a) corresponding to different model orientations are illustrated in the Table 3.7.
<table>
<thead>
<tr>
<th>Models↓</th>
<th>Input-oriented</th>
<th>Model orientation</th>
<th>Output-oriented</th>
<th>Non-oriented</th>
</tr>
</thead>
<tbody>
<tr>
<td>Super Radial (Andersen and Petersen, 1993)</td>
<td>( \min \theta^\text{super} )</td>
<td>subject to</td>
<td>( \max \phi^\text{super} )</td>
<td>subject to</td>
</tr>
<tr>
<td></td>
<td>( \sum_{j=1}^{n} \lambda_j x_{ij} + s_i^- = \theta^\text{super} x_{io} )</td>
<td>( \sum_{j=1}^{n} \lambda_j y_{ij} - s_r^- = y_{ro} )</td>
<td>( \lambda_j, s_i^-, s_r^- \geq 0 )</td>
<td>( \lambda_j, s_i^-, s_r^- \geq 0 )</td>
</tr>
<tr>
<td>Super Non-radial (Tone, 2002a)</td>
<td>( \min \rho_{\text{super}} = 1 + \frac{1}{m} \sum_{i=1}^{m} s_i^- / x_{io} )</td>
<td>subject to</td>
<td>( \min \rho_{\text{super}} = \frac{1}{1 - \frac{1}{s} \sum_{r=1}^{s} s_r^- / y_{ro}} )</td>
<td>subject to</td>
</tr>
<tr>
<td></td>
<td>( x_{io} - \sum_{j=1}^{n} \lambda_j x_{ij} + s_i^- \geq 0 )</td>
<td>( x_{io} - \sum_{j=1}^{n} \lambda_j x_{ij} - s_i^- = 0 )</td>
<td>( x_{ro} - \sum_{j=1}^{n} \lambda_j y_{ij} + s_i^- \geq 0 )</td>
<td>subject to</td>
</tr>
<tr>
<td></td>
<td>( \sum_{j=1}^{n} \lambda_j y_{ij} - y_{ro} + s_r^- = 0 )</td>
<td>( \sum_{j=1}^{n} \lambda_j y_{ij} - y_{ro} - s_i^- \geq 0 )</td>
<td>( \lambda_j, s_i^-, s_r^- \geq 0 )</td>
<td>( \lambda_j, s_i^-, s_r^- \geq 0 )</td>
</tr>
</tbody>
</table>
### 3.2.2.2 Cross-efficiency models

The cross-efficiency model was introduced by Sexton et al. (1986) and extended by Oral et al. (1991), Doyle and Green (1994) and Thanassoulis et al. (1995). This method was developed as a DEA extension tool that can be utilized to identify best-performing banks, and to rank banks using cross-efficiency scores that are linked to all banks. The basic idea of cross-efficiency models is to use DEA in a peer-appraisal instead of a self-appraisal. A peer-appraisal refers to the efficiency score of a bank that is achieved when evaluated with the optimal weights (input and output weights obtained by means of the output-oriented CRS model) of other banks. There are two principal advantages of cross-efficiency: (i) it provides a unique ordering of the banks, and (ii) it eliminates unrealistic weight schemes without requiring the elicitation of weight restrictions from application area experts (Liang et al., 2008).

To compute the mean cross-efficiency score, consider \( n \) banks that are to be evaluated in terms of \( m \) inputs and \( s \) outputs. Let \( x_{ij} (i = 1, \ldots, m) \) and \( y_{rj} (r = 1, \ldots, s) \) be the input and output values of bank \( j (j = 1, \ldots, n) \). Then, the efficiencies of the \( n \) banks can be defined as:

\[
\theta_j = \frac{\sum_{r=1}^{s} \mu_r y_{rj}}{\sum_{i=1}^{m} \nu_i x_{ij}}
\]

where \( \nu_i \) and \( \mu_r \) are input and output weights, respectively.

For a specific bank, say bank \( k, k \in \{1, \ldots, n\} \), its efficiency relative to the other banks can be measured by the following model:

\[
\begin{align*}
\max \quad & \theta_{kk} = \sum_{r=1}^{s} u_{rk} y_{rk} \\
\text{subject to} \quad & \sum_{r=1}^{s} u_{rk} y_{rj} - \sum_{i=1}^{m} \nu_{ik} x_{ik} \leq 0 \\
& \sum_{i=1}^{m} \nu_{ik} x_{ik} = 1 \\
& u_{rk}, \nu_{ik} \geq 0
\end{align*}
\]
where $v_{ik}$ and $u_{rk}$ are decision variables. Let $v_{ik}^*$ and $u_{rk}^*$ be an optimal solution to the model (3.9). Then $\theta_{ik}^* = \sum_{r=1}^s u_{rk}^* v_{ik}^*$ is referred to as the CCR-efficiency or simple efficiency of bank $k$, which is the best relative efficiency that bank $k$ can achieve, and

$$\theta_{jk} = \frac{\sum_{r=1}^s u_{rk}^* y_{rj}}{\sum_{i=1}^n v_{ik}^* x_{ik}}$$

as a cross-efficiency of bank $j$, which reflects the peer-appraisal of bank $k$ to bank $j$ ($j = 1,\ldots,n; j \neq k$). If $\theta_{ik}^* = 1$, then bank $k$ is referred to as CCR efficient; otherwise, it is referred to as inefficient. All efficient units determine an efficient frontier.

The CCR model is solved for each bank individually. As a result, there are $n$ sets of input and output weights for the $n$ banks. Each bank has $(n-1)$ cross-efficiencies plus one CCR-efficiency where $\theta_{ik}$ ($k = 1,\ldots,n$) are CCR-efficiencies of the $n$ banks, i.e., $\theta_{ik} = \theta_{ik}^*$. Since CCR model may have multiple optimal solutions, this non-uniqueness could potentially hamper the use of cross-efficiency (Baker and Talluri, 1997). To resolve this problem, Sexton et al. (1986) introduced a secondary goal to avoid the arbitrariness of cross-efficiency. One of the most commonly used secondary goals is the so-called aggressive formulation for cross-efficiency evaluation suggested by Doyle and Green (1994) which aims to minimize the cross-efficiencies of the other banks as follows:

$$\min \sum_{r=1}^s u_{rk} \left( \sum_{j=1}^n \sum_{j \neq k}^n y_{rj} \right)$$

subject to

$$\sum_{i=1}^m v_{ik} \left( \sum_{j=1}^n \sum_{j \neq k} x_{ij} \right) = 1$$

(3.10)

$$\sum_{r=1}^s u_{rk} y_{rj} - \theta_{ik}^* \sum_{i=1}^m v_{ik} x_{ik} = 0$$

$$\sum_{r=1}^s u_{rk} y_{rj} - \sum_{i=1}^m v_{ik} x_{ij} \leq 0, \quad j = 1,\ldots,n; j \neq k$$

$$u_{rk}, v_{ik} \geq 0$$

A bank is categorized as being overall efficient when it has high average cross-efficiencies, conversely, when it has lower values, it is known as ‘false
standard’ efficient bank. This can also separate the difference between the cross-efficiency score and the original technical efficiency score, and thus, the evaluated weights can be more meaningful. Once the weighting scheme and the cross-efficiencies have been found, we construct a matrix called the ‘cross-efficiencies matrix’. Such a matrix for six banks is shown in Table 3.8.

<table>
<thead>
<tr>
<th>Bank</th>
<th>Bank receiving weights</th>
<th>Average appraisal of peers</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>$\theta_{1,1}$ $\theta_{1,2}$ $\theta_{1,3}$ $\theta_{1,4}$ $\theta_{1,5}$ $\theta_{1,6}$</td>
<td>$A_1$</td>
</tr>
<tr>
<td>2</td>
<td>$\theta_{2,1}$ $\theta_{2,2}$ $\theta_{2,3}$ $\theta_{2,4}$ $\theta_{2,5}$ $\theta_{2,6}$</td>
<td>$A_2$</td>
</tr>
<tr>
<td>3</td>
<td>$\theta_{3,1}$ $\theta_{3,2}$ $\theta_{3,3}$ $\theta_{3,4}$ $\theta_{3,5}$ $\theta_{3,6}$</td>
<td>$A_3$</td>
</tr>
<tr>
<td>4</td>
<td>$\theta_{4,1}$ $\theta_{4,2}$ $\theta_{4,3}$ $\theta_{4,4}$ $\theta_{4,5}$ $\theta_{4,6}$</td>
<td>$A_4$</td>
</tr>
<tr>
<td>5</td>
<td>$\theta_{5,1}$ $\theta_{5,2}$ $\theta_{5,3}$ $\theta_{5,4}$ $\theta_{5,5}$ $\theta_{5,6}$</td>
<td>$A_5$</td>
</tr>
<tr>
<td>6</td>
<td>$\theta_{6,1}$ $\theta_{6,2}$ $\theta_{6,3}$ $\theta_{6,4}$ $\theta_{6,5}$ $\theta_{6,6}$</td>
<td>$A_6$</td>
</tr>
<tr>
<td></td>
<td>$e_1$ $e_2$ $e_3$ $e_4$ $e_5$ $e_6$</td>
<td>Average appraisal by peers</td>
</tr>
</tbody>
</table>

Here, $\theta_{jk}$ is the efficiency score calculated by the classic DEA model, $\theta_{jk}$ is the cross-efficiency score for the bank $k$ calculated using the weighting scheme obtained for the bank $j$. The leading diagonal is the special case where bank $k$ rates itself. Further, $e_k$ is the mean cross-efficiency of the bank $k$ and is calculated in the following way:

$$e_k = \frac{1}{(N-1)} \sum_{j \neq k} \theta_{j,k}$$

### 3.2.2.3 Non-discretionary input and output variables models

Banker and Morey (1986) introduced the DEA models that can be used to model non-discretionary (or uncontrollable) input and output variables. These variables are exogenously fixed variables and not under the control of bank but have significant effect on their performance (Zhu, 2003). The models incorporating non-discretionary variables are unique in the sense that (i) the radial contraction ($\theta$) in the inputs, or radial expansion ($\phi$) in the outputs, cannot be applied to the non-discretionary variables, and (ii) it eliminates the slacks for non-discretionary inputs and outputs from the objective function since the management has no control over these variables so not interested in their slacks.
To compute the relative efficiency of a bank, let us suppose that the input and output variables may each be partitioned into subsets of discretionary (D) and nondiscretionary (ND) variables. Thus,

\[ i = \{1, \ldots, m\} = i_D \cup i_{ND} \text{ with } i_D \cap i_{ND} = \emptyset \]

and

\[ r = \{1, \ldots, s\} = r_D \cup r_{ND} \text{ with } r_D \cap r_{ND} = \emptyset \]

where \( i_D \), \( r_D \), \( i_{ND} \), and \( r_{ND} \) refer to discretionary (D) and nondiscretionary (ND) input and output variables, respectively, and \( \emptyset \) is empty set. The linear programming problems of primal (multiplier) and dual (envelopment) form of the models incorporating non-discretionary input and output variables are defined as follows:

<table>
<thead>
<tr>
<th>Frontier Type</th>
<th>Non-discretionary inputs</th>
<th>Non-discretionary outputs</th>
</tr>
</thead>
</table>
| Multiplier form | \[
\begin{align*}
\max_{\nu, \mu} & \quad \sum_{j=1}^{n} \mu_j y_{rj} - \sum_{i=1}^{m} \nu_i x_{io} \\
\text{subject to} & \quad \sum_{i=1}^{m} \mu_i y_{ij} - \sum_{i=1}^{m} \nu_i x_{io} - \sum_{j=1}^{n} \nu_j x_{io} \leq 0 \\
& \quad \sum_{i=1}^{m} \nu_i x_{io} = 1 \\
& \quad \nu_i \geq \varepsilon, \quad i \in D \\
& \quad \nu_i \geq 0, \quad i \in ND \\
& \quad \mu_i \geq \varepsilon 
\end{align*}
\] | \[
\begin{align*}
\min_{\nu, \mu} & \quad \sum_{j=1}^{n} \nu_j x_{io} - \sum_{r=1}^{s} \mu_r y_{rj} \\
\text{subject to} & \quad \sum_{r=1}^{s} \mu_r y_{rj} - \sum_{r=1}^{s} \nu_j x_{io} - \sum_{r=1}^{s} \mu_r y_{rj} \geq 0 \\
& \quad \sum_{r=1}^{s} \mu_r y_{rj} = 1 \\
& \quad \nu_j \geq \varepsilon \\
& \quad \mu_r \geq \varepsilon, \quad i \in D \\
& \quad \mu_r \geq 0, \quad i \in ND 
\end{align*}
\] |
| Envelopment form | \[
\begin{align*}
\min_{\lambda, \theta, \phi} & \quad \theta - \varepsilon \left( \sum_{j=1}^{n} s_j^* + \sum_{r=1}^{s} s_r^* \right) \\
\text{subject to} & \quad \sum_{j=1}^{n} \lambda_j x_{ij} + s_i^* = \theta x_{io}, \quad i \in D \\
& \quad \sum_{j=1}^{n} \lambda_j x_{ij} + s_i^* = x_{io}, \quad i \in ND \\
& \quad \sum_{j=1}^{n} \lambda_j y_{ij} - s_r^* = y_{rj}, \quad r \in D \\
& \quad \sum_{j=1}^{n} \lambda_j y_{ij} - s_r^* = y_{rj}, \quad r \in ND \\
& \quad \lambda_i s_i^*, s_r^* \geq 0 
\end{align*}
\] | \[
\begin{align*}
\max_{\lambda, \theta, \phi} & \quad \phi + \varepsilon \left( \sum_{j=1}^{n} s_j^* + \sum_{r=1}^{s} s_r^* \right) \\
\text{subject to} & \quad \sum_{j=1}^{n} \lambda_j x_{ij} + s_i^* = x_{io} \\
& \quad \sum_{j=1}^{n} \lambda_j y_{ij} - s_r^* = \phi y_{rj}, \quad r \in D \\
& \quad \sum_{j=1}^{n} \lambda_j y_{ij} - s_r^* = y_{rj}, \quad r \in ND \\
& \quad \lambda_i s_i^*, s_r^* \geq 0 
\end{align*}
\] |

Note here that Banker and Morey formulation can also be expressed as a VRS model by adding the constraint \( \sum_{j=1}^{n} \lambda_j = 1 \).
3.2.2.4 Assurance region models

The most significant extension of DEA is the concept of assurance region (AR) models or restricted multiplier models as developed by Thompson et al. (1990), which imposes restrictions (constraints) on weights to control how much a bank can freely use the weights to become efficient. As noted, the only restriction on the multiplier DEA models is the positivity of the multipliers imposed by \( \varepsilon \), i.e., \( \varepsilon > 0 \). This flexibility is often advantageous in application of DEA methodology. However, in some situations, it can assign unreasonably low or excessively high values to the multipliers in an attempt to drive the efficiency rating for a particular bank as high as possible (Cooper et al., 2004). In the restricted multiplier models, lower and upper bounds can be established on a weight ratio of a given pair of inputs or outputs to assure that no bank can freely choose to become efficient through using excessive outputs or insufficient inputs (Ozcan, 2008). Thus, the banks will reassess their input usage and output production within given limits that are equivalent to policy or managerial restrictions. In order to impose the restrictions on input weights, the additional inequality constraints of the following form need to be incorporated into the multiplier DEA models:

\[
\alpha_i \leq \frac{v_i}{v_{io}} \leq \beta_i \quad i = 1, \ldots, m \quad (3.11)
\]

The restrictions to outputs weights can be imposed using the following formula:

\[
\delta_r \leq \frac{\mu_r}{\mu_{ro}} \leq \gamma_r \quad r = 1, \ldots, s \quad (3.12)
\]

Here, \( v_{io} \) and \( \mu_{ro} \) represent multipliers which serve as ‘numeraires’ in establishing the upper and lower bounds represented here by \( \alpha_i, \beta_i \) and by \( \delta_r, \gamma_r \) for the multipliers associated with inputs and outputs where \( \alpha_{io} = \beta_{io} = \delta_{ro} = \gamma_{ro} = 1 \).

The constraints (3.11) and (3.12) are called Assurance Regions of Type-I constraints as developed by Thompson et al. (1986). Each of these restrictions link either only input or only output weights. However, Thompson et al. (1990) defined a more precise form of AR models called AR Type-II. AR Type-II models are typically used where some relationship between the output and input concerned is to be reflected (Thanassoulis et al., 1995). Such a model imposes the restriction of a type \( \gamma_i v_i \geq u_r \). The multiplier form of DEA model is, therefore, modified to include AR constraints and resulting linear programming problem is as follows:
Table 3.10: Assurance region models

<table>
<thead>
<tr>
<th>Frontier Type↓</th>
<th>Model orientation</th>
<th>Input-oriented</th>
<th>Output-oriented</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>max $\sum_{i=1}^{m} \mu_i y_{ro} + \mu$</td>
<td>min $\sum_{i=1}^{m} v_i x_{io} + v$</td>
</tr>
<tr>
<td></td>
<td></td>
<td>subject to $\sum_{r=1}^{s} \mu_r y_{rj} - \sum_{i=1}^{m} v_i x_{ij} + \mu \leq 0$</td>
<td>subject to $\sum_{r=1}^{s} v_i x_{io} - \sum_{i=1}^{m} \mu_i y_{rj} + v \leq 0$</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$\sum_{i=1}^{m} v_i x_{io} = 1$</td>
<td>$\sum_{r=1}^{s} \mu_r y_{ro} = 1$</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$\alpha_i \leq \frac{v_i}{v_{io}} \leq \beta_i$</td>
<td>$\alpha_i \leq \frac{v_i}{v_{io}} \leq \beta_i$</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$\delta_i \leq \frac{\mu_i}{\mu_{ro}} \leq \gamma_i$</td>
<td>$\delta_i \leq \frac{\mu_i}{\mu_{ro}} \leq \gamma_i$</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$\gamma v_i \geq u_r$</td>
<td>$\gamma v_i \geq u_r$</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$\mu_i, v_i \geq 0(\varepsilon)$</td>
<td>$\mu_i, v_i \geq 0(\varepsilon)$</td>
</tr>
</tbody>
</table>

The generality of these AR constraints provides flexibility in use. Prices, util and other measures may be accommodated and so can mixtures of such concepts. Moreover, one can first examine provisional solutions that appear to be reasonably satisfactory to decision makers who cannot state the values for their preferences in an *a priori* manner.

### 3.2.3 Allocation DEA models

Allocation DEA models are used to estimate the cost, revenue and profit frontiers to obtain the respective efficiency scores corresponding to three behavioural goals to be pursued by the banks i.e., cost minimization, revenue maximization and profit maximization. It determines the efficiency scores for individual banks when information on prices of either inputs or outputs or both is given. In particular, allocation models are classified as: Cost efficiency DEA models, Revenue efficiency DEA models and Profit efficiency DEA models.

#### 3.2.3.1 Cost efficiency DEA models

Cost efficiency DEA models compute the cost efficiency measure for individual banks when information for prices of inputs is given. Let us explain the concept of cost efficiency as used in the frontier efficiency methodological framework. Cost efficiency (CE) measure provides how close a bank’s cost is to what a best-practice bank’s cost would be for producing the same bundle of
outputs (Weill, 2004). Measurement of cost efficiency requires the specification of an objective function and information on market prices of inputs. If the objective of the production unit is that of cost minimization, then a measure of cost efficiency is provided by the ratio of minimum cost to observed cost (Lovell, 1993). A methodical framework to measure cost efficiency of a bank dates back to the seminal work of Farrell (1957). In Farrell’s framework, input-oriented technical efficiency is just one component of cost efficiency, and in order to be cost efficient, a bank must first be technically efficient. However, another component of cost efficiency is input-oriented allocative efficiency (AE), which reflects the ability of the bank to choose the inputs in optimal proportions, given their respective prices. AE describes whether the bank is using the right mix of inputs in light of the relative price of each input. It should be noted that allocative efficiency is interpreted as a residual component of the cost efficiency of the bank and obtained from the ratio of cost and technical efficiency scores. It is significant to note that a measure of cost efficiency corresponds to the behavioural goals of the bank and a measure of technical efficiency ignores such goals.

An illustration of these efficiency measures as well as the way they are computed is given in Figure 3.6.

**Figure 3.6: Measurement of cost efficiency**

![Diagram](image)

In Figure 3.6, it is assumed that the bank uses two inputs, $x_1$ and $x_2$, to produce output $y$. The bank’s production frontier $y = f(x_1, x_2)$ is characterized by constant returns-to-scale, so that $1 = f(x_1/y, x_2/y)$, and the frontier is depicted by
the efficient unit isoquant $Y_o Y_o$. A bank is said to technically efficient if it is operating on $Y_o Y_o$. However, technical inefficiency relates to an individual bank’s failure to produce on $Y_o Y_o$. Hence, Bank P in the figure is technically inefficient. Thus, for Bank P, the technical inefficiency can be represented by the distance QP. As already noted, a measure of TE is the ratio of the minimum possible inputs of the bank (i.e., inputs usage on the frontier, given its observed output level) to the bank’s observed inputs. Accordingly, the level of TE for Bank P is defined by the ratio $OQ/OP$. It measures the proportion of inputs actually necessary to produce output. Allocative inefficiencies result from choosing the wrong input combinations given input prices. Now suppose that $CC'$ represents the ratio of input prices so that cost minimization point is $Q'$. Since the cost at point R is same as the cost at $Q'$, we measure the AE of the bank as $OR/OQ$, where the distance $RQ$ is the reduction in production costs which could occur if production occurs at $Q'$. Finally, the cost efficiency of the bank is defined as $OR/OP$, which can be considered a composite measure efficiency that includes both technical and allocative efficiencies. In fact, the relationship between CE, TE, and AE is expressed as:

$$CE = TE \times AE$$

$$\frac{OR}{OP} = \left(\frac{OQ}{OP}\right) \times \left(\frac{OR}{OQ}\right)$$

The frontier-based measures of cost efficiency always range between 0 and 1.

The banking efficiency literature spells two DEA models for estimating cost efficiency: (i) traditional cost efficiency (CE Type-I) model as proposed by Färe et al. (1985), and (ii) new cost efficiency (CE Type-II) model as suggested by Tone (2002b). It is worth noting here that the traditional cost efficiency model assumes input prices to be same across all banks. However, actual markets do not necessarily function under perfect competition and unit input prices might not be identical across banks. Tone (2002b) pointed out that the traditional cost efficiency model does not take account of the fact that costs can obviously be reduced by reducing the input factor prices. Therefore, the difference between traditional and new cost efficiency measures is that the former used the original inputs values in the constraints; while latter use the cost values of inputs in the constraints. The
linear programming problems for the envelopment form of CE Type-I and Type-II models are given as below:

<table>
<thead>
<tr>
<th>Table 3.11: Cost efficiency DEA models</th>
</tr>
</thead>
<tbody>
<tr>
<td>Objective</td>
</tr>
<tr>
<td><strong>Cost minimization</strong></td>
</tr>
<tr>
<td>min $\sum_{i=1}^{m} p_{i}^{o} \tilde{x}_{io}$</td>
</tr>
<tr>
<td>subject to</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
</tbody>
</table>

| **Cost efficiency** | $CE_{o} = \frac{\sum_{i=1}^{m} p_{i} x_{io}^{*}}{\sum_{i=1}^{m} p_{i} x_{i}}$ | $CE_{1} = \frac{\sum_{i=1}^{m} \tilde{x}_{io}^{*}}{\sum_{i=1}^{m} \tilde{x}_{io}}$ |

In the above models, $p_{i}^{o}$ is the unit price of $i^{th}$ input of bank ‘o’, $\tilde{x}_{io}$ is the optimal value of $i^{th}$ input of bank ‘o’, $x_{ij}$ is the actual value of the $i^{th}$ input, and $\tilde{x}_{y} = (p_{1} x_{1j}, ..., p_{g} x_{gj})^{T}$. Thus, $\tilde{x}_{io}$ is the optimal value of total cost of incurred by bank ‘o’ for $i^{th}$ input. In the CE Type-I model, the unit cost of bank ‘o’ be fixed at $p^{o}$, and the optimal mix $x_{io}^{*}$ that produces the output $y_{ro}$ is found. However, in CE Type-II model, the optimal input mix $\tilde{x}_{io}^{*}$ that produces output $y_{ro}$ can be found independently of the bank’s current unit price $p^{o}$.

3.2.3.2 Revenue efficiency DEA models

Revenue efficiency measures the change in a bank’s revenue adjusted for a random error, relative to the estimated revenue obtained from producing an output bundle as efficiently as the best-practice bank (Berger and Mester, 1997). If the objective of the bank is that of revenue maximization, then a measure of revenue efficiency is provided by the ratio of actual revenue to maximum or potential revenue. Any difference between the actual and potential revenue is attributable to either because of output-oriented technical inefficiency (producing too few outputs
of one or more outputs given the input quantities) or output-oriented allocative inefficiency (producing non-optimal combination of outputs given their prices).

The measurement of revenue efficiency in the efficiency frontier methodological framework is depicted graphically in Figure 3.7. It is assumed that the bank produces two outputs, \( y_1 \) and \( y_2 \) using the input \( x \). The production possibility curve is represented by \( TT' \). From the output-oriented framework, a bank is said to technically efficient if it is operating on \( TT' \). Therefore, banks B, C, D and E are output-oriented technically efficient, while Bank A lies below the frontier and is inefficient. In other words, the bank located at point A has a potential to increase the production levels of both outputs to point \( A' \) on the production possibility frontier. Thus, for Bank A, the output-oriented technical inefficiency can be represented by the distance \( AA' \). A measure of output-oriented TE is the ratio of the actual output to maximum outputs of the bank. Accordingly, the level of output-oriented TE for Bank A is defined by the ratio \( OA/OA' \). Now suppose that \( PP' \) represents the iso-revenue line, Bank D is deemed to be revenue efficient. Since the revenue at point F is same as the cost at D, we measure the output-oriented AE of the bank as \( OA'/OF \), where the distance \( A'F \) is the increase in revenue which could occur if production occurs at \( A' \). Finally, the revenue efficiency of the bank is defined as \( OA/OF \), which can be considered a composite measure efficiency that includes both output-oriented technical and allocative efficiencies. Further, the distance \( AF \) represents revenue inefficiency for Bank A. In fact, the relationship between RE, TE and AE is expressed as:

\[
\text{RE} = \frac{(OA/OF)}{(OA/OA') \times (OA'/OF)}
\]
Again, we have traditional revenue efficiency (RE Type-I) model and new revenue efficiency (RE Type-II) model for estimating revenue efficiency scores. As the case with cost efficiency, the difference between RE Type-I and Type-II is that Type-I model is traditional and commonly uses the original outputs values in constraints, while Type-II models use total revenue values of outputs in the constraints. The linear programming problems for traditional and new revenue efficiency models are given as follows:

### Table 3.12: Revenue efficiency DEA models

<table>
<thead>
<tr>
<th>Objective</th>
<th>Revenue Type-I</th>
<th>Revenue Type-II</th>
</tr>
</thead>
</table>
| Revenue maximization | \[
\text{max } \sum_{r=1}^{s} q^o_r y^*_r
\] subject to \[
\sum_{j=1}^{n} \lambda_j x_{ij} \leq x_{io}, \quad i = 1, \ldots, m
\] \[
\sum_{j=1}^{n} \lambda_j y_{rj} \geq y_{ro}, \quad r = 1, \ldots, s
\] \[
\lambda_j, y^*_r \geq 0, \quad j = 1, \ldots, n
\] | \[
\text{max } \sum_{r=1}^{s} y^*_r
\] subject to \[
\sum_{j=1}^{n} \lambda_j x_{ij} \leq x_{io}, \quad i = 1, \ldots, m
\] \[
\sum_{j=1}^{n} \lambda_j y_{rj} \geq y_{ro}, \quad r = 1, \ldots, s
\] \[
\lambda_j \geq 0, \quad j = 1, \ldots, n
\] |
| Revenue efficiency | \[
RE_o = \frac{\sum_{r=1}^{s} q^o_r y^*_r}{\sum_{r=1}^{s} q^o_r y^*_r}
\] | \[
RE_i = \frac{\sum_{r=1}^{s} y^*_r}{\sum_{r=1}^{s} y^*_r}
\] |

In the above models, \(q^o_r\) is the unit price of \(r\)th output of bank ‘o’, \(y^*_r\) is the optimal value of \(r\)th output for bank ‘o’, \(y_{rj}\) is the actual value of the \(r\)th output,
and \( \overline{\mathbf{y}}_q = \left( q_{1q}, y_{1q}, \ldots, q_{dq}, y_{dq} \right)^T \). Thus, \( \overline{y}_{ro} \) is the optimal value of total revenue earned by bank ‘o’ from the production of \( r^{th} \) output.

### 3.2.3.3 Profit efficiency DEA models

In the frontier efficiency measurement framework, a measure of profit efficiency assesses how close a bank comes to generating the maximum possible profit given the levels of input and output prices (quantities) and other exogenous conditions. In other words, profit efficiency improvements occur when a bank moves closer to the profit of a best-practice bank under the given conditions. It is provided by the ratio of actual profit to maximum profit. The idea of measuring profit efficiency is conceptualized in the Figure 3.8. In the figure, the curve OQ shows the production frontier. The actual input-output combination of the Bank A is \((x_A, y_A)\) shown by the point A. Therefore, the profit earned by Bank A is \( \pi = q_A y_A - p_A x_A \). The set of all \((x, y)\) through A which yield normalized profit \( \pi \) is shown by the line CD. The objective of the Bank A is to reach highest isoprofit line parallel to CD that can be attained at any point on or below the curve OQ. The highest such point on isoprofit line is reached at the point B representing the tangency of the isoprofit line EF with the production frontier. Let the optimal input-output bundle for Bank B is \((x^*, y^*)\). The intercept of this line OE equals the maximum normalized profit \( \pi^* \). Bank A achieves maximum profit when it is projected on the isoprofit curve EF (say at A^*), where maximum profits equals that of Bank B i.e., \( \pi^* = q_A y^*_A - p_A x^*_A = q_B y_B - p_B x_B \). Thus, profit efficiency for Bank A would be given by the ratio of actual to maximum profits i.e., \( PE_A = \pi/\pi^* \). Regarding the decomposition of profit efficiency, Kumbhakar and Lovell (2000) states:

“*A decomposition of profit efficiency into its constituent parts is somewhat arbitrary, depending on whether an input-oriented or an output-oriented measure of technical efficiency is used*. 

100
In the contemporary literature on banking efficiency, two measures of profit efficiency, namely standard profit efficiency and alternative profit efficiency, have been used by the researchers (see Berger et al., 1993b; Berger and Mester, 1997; Cooper et al., 2007; Maudos and Pastor, 2003). However, a consensus on the most adequate one was difficult to be achieved. These two measures differ whether or not we consider the existence of market power in the setting of output prices. The estimation of standard profit efficiency (SPE) is based on the assumptions that (i) banks maximize the profits in perfectly competitive input and output markets, and (ii) the prices of outputs and inputs are determined exogenously. Thus, the standard profit function is specified in term of input and output prices i.e., $\pi = f(p,q)$. In fact, SPE measures how close a bank is to producing the maximum possible profit given a particular level of input and output prices.

In contrast, the alternative profit efficiency (APE) developed by Humphrey and Pulley (1997) assumes the existence of imperfect competition or banks exercise a form of market power in choosing output prices. However, this market power is limited to output markets and banks remain competitive purchasers of inputs. Thus, alternative profit function is defined in terms of input prices and output quantities i.e., $\pi = f(p,y)$. In fact, APE measures how close a bank comes to earning maximum profits, given its output levels rather than its market prices. DeYoung and Hassan (1998) listed two advantages of specifying profits as a
function of output quantities rather than output prices: (i) it avoids having to measure output prices, which are not available for transactions services and fee-based outputs and can only be imperfectly constructed for loan outputs, and (ii) output quantities tend to vary across banks to a greater degree than do output prices, and as a result explain a larger portion of the variation in profits in regression analysis.

Berger and Mester (1997) noted that alternative profit frontier is preferred over the standard profit frontier when one or more of the following conditions hold: (i) there are substantial unmeasured differences in the quality of banking services; (ii) outputs are not completely variable, so that a bank cannot achieve every output scale and product mix; (iii) output markets are not perfectly competitive, so that banks have some market power over the prices they charge; and (iv) output prices are not accurately measured, so they do not provide accurate guides to opportunities to earn revenues and profits in the standard profit function.

The linear programming problems for (i) traditional standard profit efficiency (SPE Type-I) model as proposed by Färe et al. (1997), (ii) new standard profit efficiency (SPE Type-II) model as suggested by Cooper et al. (2007), and (iii) the alternative profit efficiency (APE) model as developed by Maudos and Pastor (2003) are given below:
### Table 3.13: Profit efficiency DEA models

<table>
<thead>
<tr>
<th>Objective</th>
<th>SPE Type-I</th>
<th>SPE Type-II</th>
<th>APE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Profit maximization</td>
<td>( \max \sum_{r=1}^{s} q_{yr} - \sum_{i=1}^{m} p_{ix} )</td>
<td>( \max \sum_{r=1}^{s} \tilde{y}<em>{yr} - \sum</em>{i=1}^{m} \tilde{x}_{xi} )</td>
<td>( \max \tilde{R}<em>{o} - \sum</em>{i=1}^{m} p_{i} \tilde{x}_{xi} )</td>
</tr>
<tr>
<td>Subject to</td>
<td>( \sum_{j=1}^{n} \lambda_{j} x_{ji} \leq \tilde{x}_{xi}, \quad i = 1, \ldots, m )</td>
<td>( \sum_{j=1}^{n} \lambda_{j} \tilde{x}<em>{ji} \leq \tilde{x}</em>{xi}, \quad i = 1, \ldots, m )</td>
<td>( \sum_{j=1}^{n} \lambda_{j} x_{ji} \leq \tilde{x}_{xi}, \quad i = 1, \ldots, m )</td>
</tr>
<tr>
<td></td>
<td>( \sum_{j=1}^{n} \lambda_{j} y_{jr} \geq \tilde{y}_{jr}, \quad r = 1, \ldots, s )</td>
<td>( \sum_{j=1}^{n} \lambda_{j} \tilde{y}<em>{jr} \geq \tilde{y}</em>{jr}, \quad r = 1, \ldots, s )</td>
<td>( \sum_{j=1}^{n} \lambda_{j} x_{ji} \leq \tilde{x}_{xi}, \quad i = 1, \ldots, m )</td>
</tr>
<tr>
<td></td>
<td>( x_{io} \geq \tilde{x}_{io} )</td>
<td>( \tilde{x}<em>{io} \geq \tilde{x}</em>{io} )</td>
<td>( x_{io} \geq \tilde{x}_{io} )</td>
</tr>
<tr>
<td></td>
<td>( y_{ro} \leq \tilde{y}_{ro} )</td>
<td>( \tilde{y}<em>{ro} \leq \tilde{y}</em>{ro} )</td>
<td>( R_{o} \leq \tilde{R}_{o} )</td>
</tr>
<tr>
<td></td>
<td>( \lambda_{j} \geq 0, \quad j = 1, \ldots, n )</td>
<td>( \lambda_{j} \geq 0, \quad j = 1, \ldots, n )</td>
<td>( \hat{\lambda}_{j} \geq 0, \quad j = 1, \ldots, n )</td>
</tr>
</tbody>
</table>

- Profit efficiency
  - \( SPE_{o} = \frac{\sum_{r=1}^{s} q_{yr} - \sum_{i=1}^{m} p_{ix}}{\sum_{r=1}^{s} q_{yr} - \sum_{i=1}^{m} p_{ixe}} \)
  - \( SPE_{i} = \frac{\sum_{r=1}^{s} \tilde{y}_{yr} - \sum_{i=1}^{m} \tilde{x}_{xi}}{\sum_{r=1}^{s} \tilde{y}_{yr} - \sum_{i=1}^{m} \tilde{x}_{xi}} \)
  - \( APE_{o} = \frac{\sum_{r=1}^{s} q_{yr} - \sum_{i=1}^{m} p_{ix}}{\tilde{R}_{o} - \sum_{i=1}^{m} p_{ixe}} \)
The profit efficiency scores so obtained are bounded above and have a maximum value of 1. It ranges over \((-\infty, 1]\) and equals 1 for a best-practice bank within the observed data. Profit efficiency can be negative since banks can throw away more than 100 percent of their potential profits.

3.3 Panel data DEA models

The above discussion on DEA models focused on the efficiency measurement in case of cross-section data setting. However, in general, using cross-section data DEA studies provide a snapshot of relative efficiency performance of banks for a particular year of study. Using longitudinal data or panel data, one can detect efficiency trends of banks over time and track the performance of each bank through a sequence of time periods. Two most common approaches in the DEA literature to capture the variations in efficiency over time in the panel data setting are (i) Window analysis, and (ii) Malmquist productivity index. This section outlines the distinctive features of these approaches.

3.3.1 Window analysis

Charnes et al. (1985) developed a method, firstly suggested by Klopp (1985), known as window analysis, which could be used for a panel data comprising the observations for various banks over a given period of time. In a panel data setting, window analysis performs DEA over time by using a moving average analogue, where a bank in each different period is treated as if it were a ‘different’ bank. Specifically, a bank’s performance in a particular period is contrasted with its performance in other periods in addition to the performance of the other banks.

The intrinsic advantages of window analysis are as follows. First, when the cross-section observations are small, most of them might be used in the construction of the frontier, reducing the discriminatory power of DEA (Coelli et al., 2005). Thus, the application of DEA to small samples can lead to the ‘self-identifiers’ problem (Gaganis and Pasiouras, 2009). The window analysis is often suggested as a solution to this problem. The windows are used with long panel data set in order to have a large number of sequential. One purpose of window analysis is to relieve degrees of freedom pressure when \(m+s\) (i.e., sum of inputs and outputs) is large relative to \(n\) (i.e., number of banks in a cross-section). As such, it provides a compromise between running DEA once on one large \(n \times T\) pooled panel and running DEA \(T\) times on \(T\) small cross sections (Fried et al., 2008).
Second, window analysis tracks dynamic efficiency trends through successive overlapping windows, and thus, allows for monitoring the performance over time. This may help the managers to take appropriate actions to augment the performance of the bank under consideration banks. Third, window analysis is a commonly used sensitivity analysis in DEA. It allows for an assessment of the stability of relative efficiency scores over time (Avkiran, 2006). The sensitivity in question is to that of external factors that may distort figures for a particular year and a varying group of reference units.

Suppose, there are observations for \( n \) different banks over \( T \) periods, it is treated as if there are \( n \times T \) different banks. In window analysis, the data set, with \( n \times T \) observations, is divided into a series of overlapping periods or windows, each of width \( w \) (\( w<T \)), and thus having \( n \times w \) banks. Hence, the first window has \( n \times w \) banks from period \( 1,2,\ldots,w \), the second one has \( n \times w \) banks for period \( (2,3,\ldots,w,w+1) \), and so on and the last window consists of \( n \times w \) banks for period \( T-w+1,\ldots,T \). In all, for a given set of \( n \times T \) observations, there will be \( T-w+1 \) separate windows with a size of \( n \times w \). These windows are analyzed separately. Then a moving average for each observation of banks is calculated by taking average of its scores from each window that attends. There would be \( w \) efficiency scores for each observation and the average of these scores is used as the efficiency measurements for the corresponding observation (Ozdincer and Ozyildirim, 2008).

Consider a hypothetical panel data set of 5 banks (\( n=5 \)) over six (\( T=6 \)) yearly periods. To perform the analysis using a three-year (\( w=3 \)) window, we proceed as follows. Each bank is represented as if it is a different bank for each of the three successive years in the first window (Year 1, Year 2, and Year 3), and an analysis of the 15 (\( nw = 5 \times 3 \)) banks is performed by using DEA model to obtain sharper and more realistic efficiency estimates. The window is then shifted one period, and an analysis is performed on the second three-year set (Year 2, Year 3, and Year 4) of the 15 banks. The process continues in this manner, shifting the window forward one period each time and concluding with a final (fourth) analysis of 5 banks for the last three years (Year 4, Year 5, and Year 6). Thus, one performs \( T-w+1 \) separate analysis, where each analysis examines \( n \times w \) banks. Table 3.14 depicts window analysis for aforementioned case with a three period moving window.
### Table 3.14: Format of DEA window analysis

<table>
<thead>
<tr>
<th>Bank</th>
<th>Period Window</th>
<th>Year 1</th>
<th>Year 2</th>
<th>Year 3</th>
<th>Year 4</th>
<th>Year 5</th>
<th>Year 6</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bank 1</td>
<td>Window 1</td>
<td>$\theta_{1,1}^1$</td>
<td>$\theta_{1,2}^1$</td>
<td>$\theta_{1,3}^1$</td>
<td>$\theta_{1,4}^1$</td>
<td>$\theta_{1,5}^1$</td>
<td>$\theta_{1,6}^1$</td>
</tr>
<tr>
<td></td>
<td>Window 2</td>
<td>$\theta_{2,1}^1$</td>
<td>$\theta_{2,2}^1$</td>
<td>$\theta_{2,3}^1$</td>
<td>$\theta_{2,4}^1$</td>
<td>$\theta_{2,5}^1$</td>
<td>$\theta_{2,6}^1$</td>
</tr>
<tr>
<td></td>
<td>Window 3</td>
<td>$\theta_{3,1}^1$</td>
<td>$\theta_{3,2}^1$</td>
<td>$\theta_{3,3}^1$</td>
<td>$\theta_{3,4}^1$</td>
<td>$\theta_{3,5}^1$</td>
<td>$\theta_{3,6}^1$</td>
</tr>
<tr>
<td></td>
<td>Window 4</td>
<td>$\theta_{4,1}^1$</td>
<td>$\theta_{4,2}^1$</td>
<td>$\theta_{4,3}^1$</td>
<td>$\theta_{4,4}^1$</td>
<td>$\theta_{4,5}^1$</td>
<td>$\theta_{4,6}^1$</td>
</tr>
<tr>
<td>Bank 2</td>
<td>Window 1</td>
<td>$\theta_{1,1}^2$</td>
<td>$\theta_{1,2}^2$</td>
<td>$\theta_{1,3}^2$</td>
<td>$\theta_{1,4}^2$</td>
<td>$\theta_{1,5}^2$</td>
<td>$\theta_{1,6}^2$</td>
</tr>
<tr>
<td></td>
<td>Window 2</td>
<td>$\theta_{2,1}^2$</td>
<td>$\theta_{2,2}^2$</td>
<td>$\theta_{2,3}^2$</td>
<td>$\theta_{2,4}^2$</td>
<td>$\theta_{2,5}^2$</td>
<td>$\theta_{2,6}^2$</td>
</tr>
<tr>
<td></td>
<td>Window 3</td>
<td>$\theta_{3,1}^2$</td>
<td>$\theta_{3,2}^2$</td>
<td>$\theta_{3,3}^2$</td>
<td>$\theta_{3,4}^2$</td>
<td>$\theta_{3,5}^2$</td>
<td>$\theta_{3,6}^2$</td>
</tr>
<tr>
<td></td>
<td>Window 4</td>
<td>$\theta_{4,1}^2$</td>
<td>$\theta_{4,2}^2$</td>
<td>$\theta_{4,3}^2$</td>
<td>$\theta_{4,4}^2$</td>
<td>$\theta_{4,5}^2$</td>
<td>$\theta_{4,6}^2$</td>
</tr>
<tr>
<td>Bank 3</td>
<td>Window 1</td>
<td>$\theta_{1,1}^3$</td>
<td>$\theta_{1,2}^3$</td>
<td>$\theta_{1,3}^3$</td>
<td>$\theta_{1,4}^3$</td>
<td>$\theta_{1,5}^3$</td>
<td>$\theta_{1,6}^3$</td>
</tr>
<tr>
<td></td>
<td>Window 2</td>
<td>$\theta_{2,1}^3$</td>
<td>$\theta_{2,2}^3$</td>
<td>$\theta_{2,3}^3$</td>
<td>$\theta_{2,4}^3$</td>
<td>$\theta_{2,5}^3$</td>
<td>$\theta_{2,6}^3$</td>
</tr>
<tr>
<td></td>
<td>Window 3</td>
<td>$\theta_{3,1}^3$</td>
<td>$\theta_{3,2}^3$</td>
<td>$\theta_{3,3}^3$</td>
<td>$\theta_{3,4}^3$</td>
<td>$\theta_{3,5}^3$</td>
<td>$\theta_{3,6}^3$</td>
</tr>
<tr>
<td></td>
<td>Window 4</td>
<td>$\theta_{4,1}^3$</td>
<td>$\theta_{4,2}^3$</td>
<td>$\theta_{4,3}^3$</td>
<td>$\theta_{4,4}^3$</td>
<td>$\theta_{4,5}^3$</td>
<td>$\theta_{4,6}^3$</td>
</tr>
<tr>
<td>Bank 4</td>
<td>Window 1</td>
<td>$\theta_{1,1}^4$</td>
<td>$\theta_{1,2}^4$</td>
<td>$\theta_{1,3}^4$</td>
<td>$\theta_{1,4}^4$</td>
<td>$\theta_{1,5}^4$</td>
<td>$\theta_{1,6}^4$</td>
</tr>
<tr>
<td></td>
<td>Window 2</td>
<td>$\theta_{2,1}^4$</td>
<td>$\theta_{2,2}^4$</td>
<td>$\theta_{2,3}^4$</td>
<td>$\theta_{2,4}^4$</td>
<td>$\theta_{2,5}^4$</td>
<td>$\theta_{2,6}^4$</td>
</tr>
<tr>
<td></td>
<td>Window 3</td>
<td>$\theta_{3,1}^4$</td>
<td>$\theta_{3,2}^4$</td>
<td>$\theta_{3,3}^4$</td>
<td>$\theta_{3,4}^4$</td>
<td>$\theta_{3,5}^4$</td>
<td>$\theta_{3,6}^4$</td>
</tr>
<tr>
<td></td>
<td>Window 4</td>
<td>$\theta_{4,1}^4$</td>
<td>$\theta_{4,2}^4$</td>
<td>$\theta_{4,3}^4$</td>
<td>$\theta_{4,4}^4$</td>
<td>$\theta_{4,5}^4$</td>
<td>$\theta_{4,6}^4$</td>
</tr>
<tr>
<td>Bank 5</td>
<td>Window 1</td>
<td>$\theta_{1,1}^5$</td>
<td>$\theta_{1,2}^5$</td>
<td>$\theta_{1,3}^5$</td>
<td>$\theta_{1,4}^5$</td>
<td>$\theta_{1,5}^5$</td>
<td>$\theta_{1,6}^5$</td>
</tr>
<tr>
<td></td>
<td>Window 2</td>
<td>$\theta_{2,1}^5$</td>
<td>$\theta_{2,2}^5$</td>
<td>$\theta_{2,3}^5$</td>
<td>$\theta_{2,4}^5$</td>
<td>$\theta_{2,5}^5$</td>
<td>$\theta_{2,6}^5$</td>
</tr>
<tr>
<td></td>
<td>Window 3</td>
<td>$\theta_{3,1}^5$</td>
<td>$\theta_{3,2}^5$</td>
<td>$\theta_{3,3}^5$</td>
<td>$\theta_{3,4}^5$</td>
<td>$\theta_{3,5}^5$</td>
<td>$\theta_{3,6}^5$</td>
</tr>
<tr>
<td></td>
<td>Window 4</td>
<td>$\theta_{4,1}^5$</td>
<td>$\theta_{4,2}^5$</td>
<td>$\theta_{4,3}^5$</td>
<td>$\theta_{4,4}^5$</td>
<td>$\theta_{4,5}^5$</td>
<td>$\theta_{4,6}^5$</td>
</tr>
</tbody>
</table>

**Note:** $\theta_{w,t}$ represents the relative efficiency of bank ‘$o$’ in window ‘$w$’ and period ‘$t$’.

**Source:** Author’s elaboration

The results of a window analysis as given in the Table 3.14 can be used for the identification of trends in efficiency performance, the stability of reference sets, and other possible insights. ‘Row views’ clarify efficiency performance of the banks in sample. Similar ‘Column views’ allow comparison of banks across different references sets and hence provide information on the stability of these scores as the references sets change. The utility of the table can be further extended by appending columns of summary statistics (mean, median, variance, range, etc.) for each bank to reveal the relative stability of each banks’ results.

#### 3.3.2 Malmquist productivity index (MPI)

The Malmquist productivity index, first initiated by Caves, Christensen, and Diewert (1982a, b) and further developed by Färe (1988), Färe et al. (1994a, 1994b), and others (e.g., Färe et al., 1997; Ray and Desli, 1997) has been widely...
used in the literature of productivity analyses. Although most of the desired properties can be inherited from the conventional Törnqvist (1936) index, the popularity of the MPI is attributed to the fact that TFP can be measured using distance functions without the requirement for information on prices or cost shares of factors.

DEA models used in the computation of MPI can be input-oriented or output-oriented. Consequently, the MPI can be defined from output-oriented when the inputs are fixed at their current levels or input-oriented perspective when the outputs are fixed at their current levels. The measurement of MPI from the output-oriented approach is to see how much more output has been produced, using a given level of inputs and the present state of technology, relative to what could be produced under a given reference technology using the same level of inputs. An alternative is to measure MPI from input-oriented approach by examining the reduction in input use, which is feasible given the need to produce a given level of output under a reference technology (Coelli et al., 2005). The idea of computing MPI from output- and input-oriented perspectives is described in detail as follows.

3.3.2.1 A graphical conceptualization

3.3.2.1.1 Output-oriented framework

Let us consider the Bank P which produces two outputs, $y_1$ and $y_2$, from a given level of input $x$, over two time periods: a base period, $t$, and an end period, $t+1$. In Figure 3.9, $PPC^t$ and $PPC^{t+1}$ refer to the production possibility curves for Bank P in the two time periods, respectively. Clearly, improvements in production technology have occurred (since $PPC^{t+1}$ is outside $PPC^t$) in a non-neutral way (since the shift in production possibility curves is skewed rather than parallel). Bank P’s actual production position has changed from $P_t$ to $P_{t+1}$ over the two periods. The fact that neither point lies on its associated production possibility curve indicates that the bank is technically inefficient in both time periods.
Now consider the production points of the different time periods separately. The technical efficiency of the bank (using Farrell’s output-oriented definition) in time period $t$ ($TE_t$) and time period $t+1$ ($TE_{t+1}$) is $TE_t = OP_t/OP_t'$ and $TE_{t+1} = OP_{t+1}/OP_{t+1}'$, respectively. Let us define the output distance function for period $t$ (denoted by $D_o^t(x_t, y_{t1}, y_{t2})$) where the subscript $t$ on the input and outputs denotes the quantities used in time period $t$, as the inverse of the maximum amount by which output could be increased (given the level of inputs remains constant) while still remaining within the feasible production possibility set. This is just the value measured by $TE_t$ and so $TE_t = D_o^t(x_t, y_{t1}, y_{t2})$. Similarly, it is the case that $TE_{t+1} = D_o^{t+1}(x_{t+1}, y_{t+1}, y_{2t+1})$ where $D_o^{t+1}(x_{t+1}, y_{t+1}, y_{2t+1})$ denotes the output distance function for period $t+1$ and the subscript $t+1$ on the input and output denotes the quantities used in time period $t+1$.

An examination of the way the productivity of the bank has changed over the two time periods can be approached in two ways, i.e., by using the technology in period $t$ as the reference technology, or by using the technology in period $t+1$ as the reference technology. Using the first method, the technical efficiency of the bank at point $P_{t+1}$ is measured by comparing actual output at time $t+1$ relative to the maximum that could be achieved given period $t$’s technology (i.e., $OP_{t+1}/OP_{t+1}'$ which can be denoted by $D_o^{t+1}(x_{t+1}, y_{t+1}, y_{2t+1})$), and this is compared to the technical efficiency of the bank at point $P_t$ measured by comparing actual output at time $t$ relative to the maximum that could be achieved, also given period...
t’s technology (i.e., $OP_t/OP_t$ which can be denoted by $D_o'(x_1, y_{1,1}, y_{2,1})$). A measure of the growth in productivity between the two periods using the technology of period $t$ as the reference technology is known as the Malmquist (Malmquist, 1953) output-oriented productivity index defined relative to the initial period’s technology ($M_o$) and is given as

$$
M_o^t = \frac{D_o'(x_1, y_{1,1}, y_{2,1})}{D_o'(x_1, y_{1,0}, y_{2,0})} \frac{OP_{t+1}/OP_t}{OP_{t+1}/OP_t} \tag{3.13}
$$

Using the second method, the technical efficiency of the bank at point $P_{t+1}$ is measured by comparing the output at time $t+1$ relative to the maximum that could be achieved given time $t+1$’s technology (i.e., $OP_{t+1}/OP_{t+1}$ which can be denoted by $D_o'^{t+1}(x_1, y_{1,1}, y_{2,1})$), and this is compared to the technical efficiency of the bank at point $P_t$ measured by comparing the output at time $t$ relative to the maximum that could be achieved also given period $t+1$’s technology (i.e., $OP_t/OP_t$ which can be denoted by $D_o(t)(x_1, y_{1,0}, y_{2,0})$). The measure of the growth in productivity between the two periods using the technology of period $t+1$ as the reference technology is known as the Malmquist output-oriented productivity index defined relative to the final period’s technology ($M_o^{t+1}$) and is given by

$$
M_o^{t+1} = \frac{D_o'^{t+1}(x_1, y_{1,1}, y_{2,1})}{D_o'(x_1, y_{1,0}, y_{2,0})} \frac{OP_{t+1}/OP_t}{OP_{t+1}/OP_t} \tag{3.14}
$$

We, therefore, have two measures of the change in productivity over the two periods ($n$ measures in the $n$ period case) and it is unclear which measure is the appropriate one to use, since the choice of base technology would be arbitrary. This problem is overcome by using the Malmquist output-oriented productivity change index ($M_o$) which is defined as the geometric mean of $M_o'$ and $M_o^{t+1}$ (Färe et al., 1994a):

$$
M_o = \sqrt{M_o^t \cdot M_o^{t+1}} = \sqrt{\frac{D_o'(x_1, y_{1,1}, y_{2,1})}{D_o'(x_1, y_{1,0}, y_{2,0})} \cdot \frac{D_o'^{t+1}(x_1, y_{1,1}, y_{2,1})}{D_o'(x_1, y_{1,0}, y_{2,0})} \frac{OP_{t+1}/OP_t}{OP_{t+1}/OP_t} \cdot \frac{OP_{t+1}/OP_t}{OP_{t+1}/OP_t}} \tag{3.15}
$$

The index (3.15) can be rewritten as:

109
The first component of equation (3.16) is the ratio of technical efficiency in time period $t+1$ (the final period) to technical efficiency in time period $t$ (the initial period), and therefore, measures the change in technical efficiency between the two periods. The ratio equals 1 if there is no change in technical efficiency over the two periods, and is greater than 1 (less than 1) if technical efficiency has improved (declined) over the two periods.

The second component measures the change in production technology (i.e., shifts in the frontier) between the two periods $t$ and $t+1$. It is the geometric mean of the change in technology between the two periods evaluated at $x_t$ and $x_{t+1}$, respectively. This component has the value 1 when there is no change in production technology, and is greater than 1 (less than 1) if change in production technology has had a positive (negative) effect.

### 3.3.2.1.2 Input-oriented framework

The Malmquist productivity index can also be defined in an input-oriented framework. First, we must define the input distance function which is the maximum amount by which all inputs could be reduced (given the level of outputs remains constant) while still remaining in the feasible input set. In Figure 3.10, the bank now uses two inputs $x_1$ and $x_2$ to produce output $y$, and $I_t$ and $I_{t+1}$ refer to the isoquants in the period $t$ and $t+1$, respectively. Improvements in production technology have occurred, since $I_{t+1}$ is inside $I_t$. The observed production points for the bank in time periods $t$ and $t+1$, are $P_t$ and $P_{t+1}$, respectively, neither of which is technically efficient since each lies beyond its own isoquant. The value of the input distance function for the bank in time period $t$ is $D'_i(x_{it}, x_{2t}, y_t) = OP_t / OP'_t$. This is the reciprocal of the Farrell input-oriented measure of technical efficiency for Bank P at time $t$. Similarly, the distance function for Bank P in time period $t+1$ is $D''_i(x_{it+1}, x_{2t+1}, y_{t+1}) = OP_{t+1} / OP''_{t+1}$.
As with the output-oriented approach, the measurement of how productivity, in input-oriented context, has changed over two time periods can be approached in two ways, i.e., by using period $t$ technology, or by using period $t+1$ technology. Using the first method, the technical inefficiency of the bank at point $P_{t+1}$ is measured by comparing the actual input at time $t+1$ relative to the minimum required given period $t$’s technology (i.e., $OP_{t+1} / OP_t$, which can be denoted by $D_t^t(x_{1t+1}, x_{2t+1}, y_{1t+1})$). This is compared to the technical inefficiency of the bank at point $P_t$ measured by comparing the actual input at time $t$ relative to the minimum required, also given period $t$’s technology (i.e., $OP_t / OP_t^t = D_t^t(x_{1t}, x_{2t}, y_{1t})$). A measure of the change in productivity between the two periods given the technology of period $t$ as the reference technology is known as the Malmquist input-oriented productivity index defined relative to the initial period’s technology ($M_t^t$) and is given as

$$M_t^t = \frac{D_t^t(x_{1t+1}, x_{2t+1}, y_{1t+1})}{D_t^t(x_{1t}, x_{2t}, y_{1t})} = \frac{OP_{t+1}}{OP_t} / \frac{OP_{t+1}^t}{OP_t^t} \quad (3.17)$$

Using the second method, the technical inefficiency of the bank at point $P_{t+1}$ is measured by comparing the actual input at time $t$ relative to the minimum input required given the technology of period $t+1$ (i.e., $OP_{t+1} / OP_{t+1}^t = D_t^{t+1}(x_{1t+1}, x_{2t+1}, y_{1t+1})$). This is compared with the technical inefficiency of the bank at point $P_t$ measured by comparing the actual input at time
the minimum input required given period $t+1$’s technology (i.e., $OP_t/OP_t^{t+1}$ which can be denoted by $D^t_i(x_{t1},x_{t2},y_t)$). The Malmquist input-oriented productivity index defined relative to the final period’s technology ($M_i^{t+1}$) is given as

$$M_i^{t+1} = \frac{D^t_i(x_{t1},x_{t2},y_t)}{D^{t+1}_i(x_{t1},x_{t2},y_t)} = \frac{OP_t/OP_t^{t+1}}{OP_t/OP_t^{t+1}}$$  \hspace{1cm} (3.18)

The problem of the arbitrary choice of which technology to use as the base technology when comparing productivity change over two periods is again overcome by defining the Malmquist input-oriented productivity change index (denoted by $M_i$) as the geometric mean of $M'_i$ and $M_i^{t+1}$:

$$M_i = \sqrt{M'_i \cdot M_i^{t+1}} = \sqrt{\frac{D^t_i(x_{t1},x_{t2},y_t)}{D^{t+1}_i(x_{t1},x_{t2},y_t)} \cdot \frac{D_i^{t+1}(x_{t1},x_{t2},y_{t+1})}{D_i^t(x_{t1},x_{t2},y_t)}} = \sqrt{\frac{OP_t/OP_t^{t+1}}{OP_t/OP_t^{t+1}} \cdot \frac{OP_t^{t+1}/OP_t^{t+1}}{OP_t^{t+1}/OP_t^{t+1}}}}$$  \hspace{1cm} (3.19)

This can be rewritten as

$$M_i = \frac{D^t_i(x_{t1},x_{t2},y_t)}{D^{t+1}_i(x_{t1},x_{t2},y_t)} \cdot \sqrt{\frac{D_i^{t+1}(x_{t1},x_{t2},y_{t+1})}{D_i^t(x_{t1},x_{t2},y_t)} \cdot \frac{D_i^t(x_{t1},x_{t2},y_t)}{D_i^{t+1}(x_{t1},x_{t2},y_{t+1})}} = \frac{OP_t/OP_t^{t+1}}{OP_t/OP_t^{t+1}} \cdot \sqrt{\frac{OP_t^{t+1}/OP_t^{t+1}}{OP_t^{t+1}/OP_t^{t+1}} \cdot \frac{OP_t/OP_t^{t+1}}{OP_t/OP_t^{t+1}}}$$  \hspace{1cm} (3.20)

The components of this index can be interpreted in the opposite way from components of the output-oriented productivity index of equation (3.16). Specifically, $D_i^{t+1}(x_{t1},x_{t2},y_{t+1})/D_i^t(x_{t1},x_{t2},y_t)$ represents the change in technical efficiency between periods $t+1$ and $t$ and equals 1 if there has been no change, and is less than 1 (greater than 1) if there has been an improvement (decline) in technical efficiency. The second component, measures the change in production technology between periods $t+1$ and $t$ and equals 1 if there has been no change, and is less than 1 (greater than 1) if the effects of production technology have been positive (negative).

3.3.2.2 DEA-based estimation of Malmquist productivity index

As noted above, the MPI productivity index can be calculated either by using input-oriented and output-oriented distance functions. Distance functions can
represent a multi input-multi output technology without any behavioral assumptions such as cost minimization or profit maximization.

Let \( x^t = (x_1^t, \ldots, x_m^t) \) denote a vector of \( m \) inputs at time \( t \) and \( y^t = (y_1^t, \ldots, y_s^t) \) be a vector of \( s \) outputs at time \( t \). The production technology at time \( t \), \( T^t \) is defined by

\[
T^t = \left\{ (x', y') : x' \text{ can produce } y' \right\}
\]

and it consists of all input-output vectors that are technically feasible at time \( t \). The Shephard’s (1970) input distance function is defined on the technology \( T \) as

\[
D_i^t(y', x') = \sup \left\{ \theta : \left( \frac{x'}{\theta}, y' \right) \in T^t \right\}
\]

e.i., as the ‘maximal’ feasible contraction of \( x' \).

The output distance function due to Shephard (1970) is defined by

\[
D_o^t(x', y') = \inf \left\{ \phi : \left( \frac{x'}{\phi}, y' \right) \in T^t \right\}
\]

e.i., the ‘minimal’ feasible expansion of \( y' \) (Fung et al., 2008).

Assume two time periods \( t \) and \( t+1 \), respectively, and define in each one of them technology and production as above. Taking time period \( t \) as the reference period, the input- and output-oriented MPI are given as:

\[
MPI_i^{t+1}(y'^{t+1}, x'^{t+1}, y', x') = \frac{D_i^{t+1}(y'^{t+1}, x'^{t+1} | CRS)}{D_i^{t}(y', x' | CRS)} \times \frac{D_o^{t+1}(y'^{t+1}, x'^{t+1} | CRS)}{D_o^{t}(y', x' | CRS)}
\]

\[
 MPI_o^{t+1}(x'^{t+1}, y'^{t+1}, x', y') = \frac{D_o^{t+1}(x'^{t+1}, y'^{t+1} | CRS)}{D_o^{t}(x', y' | CRS)} \times \frac{D_i^{t+1}(x'^{t+1}, y'^{t+1} | CRS)}{D_i^{t}(x', y' | CRS)}
\]

(3.21)

and

(3.22)

The ‘CRS’ stands for constant returns-to-scale, and it is explicitly recognize that the distance functions are defined relative to CRS technology. From equations (3.21) and (3.22), we note that the MPI is thus defined as the product of efficiency change (EFFCH), which is how much closer a bank gets to the efficient frontier (catching-up effect or falling behind), and technical change (TECH), which is how much the benchmark production frontier shifts at each bank’s observed input mix.
(technical progress or regress). MPI can attain a value greater than, equal to, or less than unity depending on whether the bank experiences productivity growth, stagnation or productivity decline, respectively, between periods $t$ and $t+1$. Similarly, EFFCH index takes a value greater than 1 for an efficiency increase, 0 for no efficiency change, or less than 1 for an efficiency decrease. Likewise, TECH attains a value greater than 1 for technical progress, 0 for technical stagnation, or less than 1 for technical regress.

In order to calculate the productivity of bank ‘$o$’ between $t$ and $t + 1$, we need to solve four different distance functions that make up MPI using either radial or non-radial DEA models. It is worth noting here that both radial and non-radial models compute a distance function as a reciprocal of Farrell’s (1957) measure of technical efficiency. For radial MPI, we should consider the linear programming problems outlined below in Table 3.15.
Table 3.15: Linear programming problems for radial input- and output-oriented distance functions

<table>
<thead>
<tr>
<th>Panel</th>
<th>Distance function at time $t$ using the reference technology for the period $t$</th>
<th>Distance function at time $t+1$ using the reference technology for the period $t+1$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Panel A:</td>
<td>$\begin{bmatrix} \tilde{D}_i (y'<em>o, x'<em>o) \end{bmatrix} = \min</em>{\theta_o, \lambda} \theta_o$ subject to $\sum</em>{j}^{n} \lambda_j x'_o \leq \theta_o x'<em>o$ $\sum</em>{j}^{n} \lambda_j y'_o \geq y'_o$ $\lambda_j \geq 0$</td>
<td>$\begin{bmatrix} \tilde{D}<em>i (x'<em>o) \end{bmatrix} = \max</em>{\phi_o} \phi_o$ subject to $\sum</em>{j}^{n} \lambda_j x'_o \leq x'<em>o$ $\sum</em>{j}^{n} \lambda_j y'_o \geq \phi_o y'_o$ $\lambda_j \geq 0$</td>
</tr>
<tr>
<td>Panel B:</td>
<td>$\begin{bmatrix} \tilde{D}<em>{i+1} (y'</em>{o}, x'<em>{o}) \end{bmatrix} = \min</em>{\theta_o, \lambda} \theta_o$ subject to $\sum_{j}^{n} \lambda_j x'<em>{o+j} \leq \theta_o x'</em>{o}$ $\sum_{j}^{n} \lambda_j y'<em>{o+j} \geq y'</em>{o}$ $\lambda_j \geq 0$</td>
<td>$\begin{bmatrix} \tilde{D}<em>{i+1} (x'</em>{o}, y'<em>{o}) \end{bmatrix} = \max</em>{\phi_o} \phi_o$ subject to $\sum_{j}^{n} \lambda_j x'<em>{o+j} \leq x'</em>{o}$ $\sum_{j}^{n} \lambda_j y'<em>{o+j} \geq \phi_o y'</em>{o}$ $\lambda_j \geq 0$</td>
</tr>
<tr>
<td>Panel C:</td>
<td>$\begin{bmatrix} \tilde{D}<em>i (y'</em>{o}, x'<em>{o}) \end{bmatrix} = \min</em>{\theta_o, \lambda} \theta_o$ subject to $\sum_{j}^{n} \lambda_j x'<em>{o+j} \leq \theta_o x'</em>{o}$ $\sum_{j}^{n} \lambda_j y'<em>{o+j} \geq y'</em>{o}$ $\lambda_j \geq 0$</td>
<td>$\begin{bmatrix} \tilde{D}<em>i (x'</em>{o}, y'<em>{o}) \end{bmatrix} = \max</em>{\phi_o} \phi_o$ subject to $\sum_{j}^{n} \lambda_j x'<em>{o+j} \leq x'</em>{o}$ $\sum_{j}^{n} \lambda_j y'<em>{o+j} \geq \phi_o y'</em>{o}$ $\lambda_j \geq 0$</td>
</tr>
<tr>
<td>Panel D:</td>
<td>$\begin{bmatrix} \tilde{D}<em>i (y'</em>{o}, x'<em>{o}) \end{bmatrix} = \min</em>{\theta_o, \lambda} \theta_o$ subject to $\sum_{j}^{n} \lambda_j x'<em>{o+j} \leq \theta_o x'</em>{o}$ $\sum_{j}^{n} \lambda_j y'<em>{o+j} \geq y'</em>{o}$ $\lambda_j \geq 0$</td>
<td>$\begin{bmatrix} \tilde{D}<em>i (x'</em>{o}, y'<em>{o}) \end{bmatrix} = \max</em>{\phi_o} \phi_o$ subject to $\sum_{j}^{n} \lambda_j x'<em>{o+j} \leq x'</em>{o}$ $\sum_{j}^{n} \lambda_j y'<em>{o+j} \geq \phi_o y'</em>{o}$ $\lambda_j \geq 0$</td>
</tr>
</tbody>
</table>

Source: Author’s elaboration
It is worth noting here that the estimation of distance functions in the radial MPI is based on DEA model developed by Charnes et al. (1978) which takes no account of slacks. However, non-zero input and output slacks are very likely to present even after the radial efficiency improvement, and often, these non-zero slack values represent a substantial amount of inefficiency. Therefore, in order to fully measure the inefficiency in bank’s performance, it is very important to also consider the inefficiency represented by the non-zero slacks in the DEA-based MPI. Tone (2001) proposed three different variants of a non-radial MPI based on slacks-based measure (SBM) model that assesses the performance of banks by simultaneously dealing with input excesses and output shortfalls of the banks concerned. The linear programming problems for estimating distance functions using SBM models are outlined in Table 3.16.
Table 3.16: Linear programming problems for non-radial (SBM) input- and output-oriented distance functions

<table>
<thead>
<tr>
<th>Panel A: Distance function at time $t$ using the reference technology for the period $t$</th>
<th>Panel B: Distance function at time $t+1$ using the reference technology for the period $t+1$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\left[ \hat{D}<em>i(y_i', x_r') \right]^{-1} = \min</em>{\vartheta, \gamma_i} 1 - \frac{1}{m} \sum_{i=1}^{m} s_i^- / x_{io}'$</td>
<td>$\left[ \hat{D}<em>i(y_i', x_r') \right]^{-1} = \min</em>{\vartheta, \gamma_i} 1 - \frac{1}{m} \sum_{i=1}^{m} s_i^- / x_{io}'$</td>
</tr>
<tr>
<td>subject to</td>
<td>subject to</td>
</tr>
<tr>
<td>$\sum_{j=1}^{n} \lambda_j x_{ij}' + s_i'^- = x_{io}'$</td>
<td>$\sum_{j=1}^{n} \lambda_j x_{ij}' \leq x_{io}'$</td>
</tr>
<tr>
<td>$\sum_{j=1}^{n} \lambda_j y_{ij}' \geq y_{ro}'$</td>
<td>$\sum_{j=1}^{n} \lambda_j y_{ij}' - s_i'^+ = y_{ro}'$</td>
</tr>
<tr>
<td>$\lambda_j, s_i'^- \geq 0$</td>
<td>$\lambda_j, s_i'^+ \geq 0$</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Panel C: Distance function at time $t$ using the reference technology for the period $t+1$</th>
<th>Panel D: Distance function at time $t+1$ using the reference technology for the period $t$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\left[ \hat{D}<em>i(y_i'^{t+1}, x_r') \right]^{-1} = \min</em>{\vartheta, \gamma_i} 1 - \frac{1}{m} \sum_{i=1}^{m} s_i^- / x_{io}'$</td>
<td>$\left[ \hat{D}<em>i(y_i'^{t+1}, x_r') \right]^{-1} = \min</em>{\vartheta, \gamma_i} 1 - \frac{1}{m} \sum_{i=1}^{m} s_i^- / x_{io}'$</td>
</tr>
<tr>
<td>subject to</td>
<td>subject to</td>
</tr>
<tr>
<td>$\sum_{j=1}^{n} \lambda_j x_{ij}' + s_i'^- = x_{io}'$</td>
<td>$\sum_{j=1}^{n} \lambda_j x_{ij}' \leq x_{io}'$</td>
</tr>
<tr>
<td>$\sum_{j=1}^{n} \lambda_j y_{ij}' \geq y_{ro}'$</td>
<td>$\sum_{j=1}^{n} \lambda_j y_{ij}' - s_i'^+ = y_{ro}'$</td>
</tr>
<tr>
<td>$\lambda_j, s_i'^- \geq 0$</td>
<td>$\lambda_j, s_i'^+ \geq 0$</td>
</tr>
</tbody>
</table>

Source: Author’s elaboration
Färe et al. (1994b) enhanced the aforementioned decomposition of MPI by taking EFFCH component, and decomposing it into pure technical efficiency change (PECH) and scale efficiency change (SECH) components with respect to variable returns-to-scale (VRS) technology. The input-oriented MPI with enhanced decomposition as developed by Färe et al. is given as:

\[
MPI^{\text{eff}}_{t+1}(y^t, x^t, y', x') = \frac{D_t\left(y^t, x^t \mid VRS\right)}{D_t\left(y', x' \mid VRS\right)} \times \frac{D_t\left(y^t, x^t \mid VRS\right) / D_t\left(y', x' \mid CRS\right)}{D_t\left(y', x' \mid VRS\right) / D_t\left(y', x' \mid CRS\right)} \times \frac{D_t\left(y^t, x^t \mid CRS\right) / D_t\left(y', x' \mid CRS\right)}{D_t\left(y', x' \mid CRS\right) / D_t\left(y', x' \mid CRS\right)}
\]

(3.23)

The corresponding output-oriented MPI decomposition can be defined in an analogous manner.

The extended decomposition of MPI by Färe et al. (1994b) has been criticized by Ray and Desli (1997) who are having the opinion that ‘their use of CRS and VRS within the same decomposition of the MPI raises a problem of internal inconsistency’. The MPI approach gives the correct estimation of TFP change in the presence of CRS technology, while may not accurately measure productivity changes when VRS is assumed for the technology (Grifell-Tatjé and Lovell, 1995). The fundamental problem is that the imposition of a VRS technology creates a systematic bias on the productivity measurement derived unless the VRS technology is identical to CRS technology (Odeck, 2008). However, various alternatives have been proposed, but none of them has gained widespread acceptance. The debate continues on how a proper Malmquist index can be derived assuming VRS, and complete redo is yet to emerge (see Grifell-Tatjé and Lovell, 1999; Balk, 2001 for discussion on this issue).

3.4 Strengths, limitations, basic requirements and outcomes of DEA

3.4.1 Strengths and limitations

Since the publication of the seminal paper of Charnes et al. in 1978, there have been thousands of theoretical contributions and practical applications in various fields using DEA (Klimberg and Ratick, 2008). The bibliographies compiled by Tavares (2002) and Emrouznejad et al. (2008) highlight that over the years, DEA has been applied in many diverse areas such as health care, banking, education, software production, telecommunication, transport, military operations,
criminal courts, electric utilities, library services, mining operations, manufacturing, etc. Some notable advantages of DEA which motivated the researchers, including us, to use it over other frontier efficiency measurement techniques are as follows.

First, DEA is able to manage complex production environments with multiple inputs and outputs technologies (Jacobs, 2000). Second, DEA optimizes for each individual observation, in place of the overall aggregation, and single optimization thereafter, performed in statistical regressions. Instead of trying to fit a regression plane through the center of the data, DEA floats a piecewise linear surface to rest on top of observations (Majumdar and Chang, 1996). Third, DEA approach has good asymptotic statistical properties. For instance, DEA is equivalent to maximum likelihood estimation, with the specification of the production frontier in DEA as a non-parametric monotone and concave function instead of a parametric form linear in parameters (Banker, 1993). Fourth, DEA produces a scalar measure of efficiency for each unit, which makes the comparison easy (Sowlati and Paradi, 2004). Fifth, in DEA, the computations are value-free and do not require specification or knowledge of a priori weights of prices for inputs or outputs (Charnes et al., 1994). DEA does not require any pre-specified functional form between inputs and outputs i.e., production function (Mirmirani et al., 2008). Therefore, the probability of a misspecification of the production technology is zero (Jemric and Vujicic, 2002). Thus, DEA estimates are unbiased if we assume that there is no underlying model or reference technology (Kittelsen, 1999). Sixth, it does not require the establishment of arbitrary cutoff points for classifying efficient and inefficient banks (Rutledge et al., 1995).

On the other hand, the main disadvantages of DEA as summarized by Coelli et al. (2005) are as follows. First, DEA does not account for random error. The deviations from the frontier are assumed to be due to inefficiency. Errors in measurement and random noise can misrepresent real relative efficiency. Second, there is a strong influence of estimated frontier by outliers. If the outlier is a high performance unit with the same characteristics as other units, then it provides a good benchmark for inefficient units. However, if the outlier is operating in a different cultural environment or has some other unique aspect, and therefore has an unfair or unattainable advantage, then other units will receive artificially low scores. Third, DEA is intended for estimating the relative efficiency of a bank, but
does not specifically address absolute efficiency. In other words, it measures how well the bank is performing compared to its peers (set of efficient units), but not compared to a theoretical maximum. The main problem arising from this is the impossibility of ranking efficient units; indeed all the efficient units have an efficiency score of 100%. Fourth, the technique requires a minimum number of units in order to guarantee the necessary degrees of freedom in the model. In principle, all inputs and outputs relevant to the function of the units should be included. However, the larger the number of inputs and outputs in relation to the number of units being assessed, the less discriminating the method appears to be (Thanassoulis, 2001). Further, the analyses containing less than the minimum number of units will yield higher efficiency scores and more units on the efficiency frontier, and hence give a more favourable picture.

3.4.2 Basic requirements

The implementation of DEA for assessing the performance of banks requires the fulfillment of certain conditions concerning (number of DMUs, the weights, the environment in which bank is operating, etc.). Ho (2004) described the following basic requirements for implementing DEA so as to get robust efficiency estimates:

(i) The banks must operate in the same cultural environment$^5$,

(ii) The model must contain suitable inputs and outputs. For example, a bank measuring productivity should employ a model with variables such as number of employees, number of branches, physical assets, etc. as opposed to square footage, etc., although these can be incorporated into the model if that is required,

(iii) Each bank must have a complete set of accurate data for all variables in the model, and

(iv) There must be a minimum number of units to study in order to maintain sufficient degrees of freedom. Cooper et al. (2007) provides two such rules to determine the minimum number of banks ($n$) that can be expressed as: $N \geq \max \{ m \times s; 3(m + s) \}$, where $N$=number of banks, $m$=number of inputs, and $s$=number of outputs. The first rule of thumb states that sample size should be greater than equal to product of inputs and outputs. While the second rule states that number of observation in the data set should be at
least three times the sum of number of input and output variables. For example, if a model includes 5 inputs and 2 output variables, there should be at least \[ \max\{5 \times 2; 3(5 + 2)\} = \max\{10, 21\} = 21 \text{ banks}. \]

3.4.3 Outcomes

An implementation of DEA gives the outcomes that not only limited to the efficiency scores alone but also related to the peers to be emulated, presence of slacks in inputs and outputs, and potential improvements in the production process. Table 3.17 highlights the implications of various outcomes of DEA for a decision making unit. The most immediate output of DEA is a list of relatively efficient and inefficient banks. Using DEA, the comparatively inefficient banks can be ranked in order of their relative inefficiency to indicate which banks are more inefficient in relation to others. An additional outcome of DEA can be found by calculating the number of times an efficient bank appears within various peer comparison sets for relatively inefficient units.

<table>
<thead>
<tr>
<th>S. No.</th>
<th>DEA outcomes</th>
<th>Implications</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>Overall efficiency score (0-100%)</td>
<td>a) Quantitative ranking of the banks in an objective, fair, strict and unambiguous manner, and b) Gauging the extent of inefficiency.</td>
</tr>
<tr>
<td>2.</td>
<td>Peer analysis</td>
<td>Identify best-performer(s) in the sample whose practices can be emulated by inefficient bank(s) in the sample.</td>
</tr>
<tr>
<td>3.</td>
<td>Slacks</td>
<td>a) Helps to identify waste of critical inputs, b) Helps to identify overproduction of output, and c) Helps in designing and implementation of an efficiency improvement program.</td>
</tr>
<tr>
<td>4.</td>
<td>Targets</td>
<td>Potential reduction in inputs and potential augmentation of outputs for projecting inefficient banks on to the efficient frontier. The targets also help in designing and implementation of an efficiency improvement program.</td>
</tr>
</tbody>
</table>

Source: Author’s elaboration

3.5 Free disposal hull analysis (FDH)

The free disposal hull approach (FDH) was proposed by Deprins et al. (1984) and extended by Tulkens (1993). It is a more general version of the DEA variable returns-to-scale model as it relies only on the strong free disposability assumption for production set, and hence does not restrict itself to convex technologies. If the true production set is convex then the DEA and FDH are both consistent estimators. However, FDH shows a lower rate of convergence (due to the less assumption it require) with respect to DEA. On the contrary, if the true
production set is not convex, than DEA is not a consistent estimator of the production set, while FDH is consistent.

As deterministic non-parametric methods, DEA and FDH assume no particular functional form for the boundary and ignore measurement error. Instead, the best-practice technology is the boundary of a reconstructed production possibility subset based upon directly enveloping a set of observations. However, in FDH, the production possibilities set is composed only of the DEA vertices and the free disposal hull points interior to these vertices (Berger and Humphrey, 1997). Since the DEA presumes that linear substitution is possible between observed input combinations on an isoquant which is generated from the observations in piecewise linear form, FDH presumes that no substitution is possible so the isoquant looks like a step function formed by the intersection of lines drawn from observed (local) Leontief-type input combinations. Thus, the FDH frontier has a staircase shape and envelops the data more tightly than the DEA frontier does. Moreover, the FDH frontier is either congruent with or interior to the DEA frontier, and thus, FDH will typically generate larger estimates of average efficiency than DEA (Tulkens, 1993).

Figure 3.11 compares DEA and FDH frontiers. The production frontier of FDH as represented by a staircase line ABCDEF is contained in DEA-BCC frontier ABCEF, which in turn is contained in DEA-CCR frontier OCG. The DEA-CCR assumes constant returns-to-scale so that all observed production combinations can be scaled up or down proportionally, and DEA-BCC assumes variable returns-to-scale and is represented by a piecewise linear convex frontier. Thus, FDH, DEA-CCR and DEA-BCC models define different production possibility sets and efficiency results. As an example, the input-oriented efficiency of Bank T is given by \( OI/OT \), \( OH/OT \) and \( OG/OT \) as determined by the FDH, DEA-CCR and DEA-BCC models, respectively.
For computing efficiency measures, the FDH model is formulated by adding the additional constraint $\lambda_j \in \{0,1\}$ i.e., $\lambda_j$ to be binary in the DEA-BCC model so as to relax an assumption of convexity.

$$\min \theta^{FDH}$$

subject to

$$\sum_{j=1}^{n} \lambda_j x_{ij} \leq \theta^{FDH} x_{i0} \quad i = 1, \ldots, m$$

$$\sum_{j=1}^{n} \lambda_j y_{rj} \geq y_{ro} \quad r = 1, \ldots, s$$

$$\sum_{j=1}^{n} \lambda_j = 1 \quad j = 1, \ldots, n$$

$$\lambda_j \in \{0,1\} \ \text{i.e., binary}$$

**3.6 Stochastic frontier analysis (SFA)**

Another widely used frontier efficiency methodological framework used in applied banking efficiency analyses is stochastic frontier analysis (SFA) which was independently proposed by Aigner et al. (1977) and Meeusen and van den Broeck (1977). Kumbhakar and Lovell (2000) in their excellent book provided an extensive survey of literature on SFA, and therefore we restrict ourselves on the broad contours of this frontier efficiency measurement approach. Coelli et al. (2003) listed out the main advantages of SFA methodology as: (i) environment variables are easier to deal with, (ii) it allows us to conduct statistical tests of hypotheses concerning any parameter restrictions associated with economic theory,
and (iii) allows an easier identification of outliers. On the other hand, the main disadvantages of SFA are that estimation results are sensitive to distributional assumptions on the error terms, and it requires large samples for robustness, so it is very demanding regarding data requirements.

The stochastic frontier involves two random components, one associated with the presence of technical inefficiency and the other being a traditional random error. An appropriate formulation of a stochastic frontier model in terms of a general production function for the $i^{th}$ bank is

$$y_i = f(x_i; \beta) e^{v_i} e^{u_i}$$

(3.25)

where $y_i$ indicates the observed level of output, $x_i$ is the vector of inputs used in the production process, $\beta$ is the vector of unknown technological parameters to be estimated, $f(x_i; \beta)$ is the deterministic part of the production function, $v_i$ is the two-sided ‘white noise’ component representing random disturbance, and $u_i$ is a non-negative one-sided disturbance term which accounts for technical inefficiency.

In logarithmic terms, this can be written as

$$\ln y_i = \ln f(x_i; \beta) + v_i - u_i = \ln f(x_i; \beta) + \epsilon_i$$

(3.26)

The stochastic production frontier model (3.26) is often referred to as a ‘composed error’ model since the error term has two components. The important features of the stochastic frontier model can be illustrated graphically. Let us suppose the bank produces the output $y_i$ using only one input $x_i$. In this case, a Cobb-Douglas stochastic frontier model takes the form:

$$\ln y_i = \beta_0 + \beta_1 \ln x_i + v_i - u_i$$

or

$$y_i = \exp(\beta_0 + \beta_1 \ln x_i + v_i - u_i)$$

(3.27)

or

$$y_i = \exp(\beta_0 + \beta_1 \ln x_i) \times \exp(v_i) \times \exp(-u_i)$$

such a frontier is depicted in Figure 3.12 where we plot inputs and outputs of two banks, A and B, and where the deterministic component of the frontier model has been drawn to reflect the existence of diminishing returns-to-scale. Values of the input are measured along the horizontal axis and outputs are measured on the vertical axis. Bank A uses the input level $x_A$ to produce the output $y_A$, while Bank B uses the input level $x_B$ to produce the output $y_B$ (these observed values are indicated by the points marked with ×). If there were no inefficiency
effects (i.e., if \( u_A = 0 \) and \( u_B = 0 \)) then so-called frontier outputs would be
\[
y^*_A \equiv \exp(\beta_0 + \beta_1 \ln x_A + v_A) \quad \text{and} \quad y^*_B \equiv \exp(\beta_0 + \beta_1 \ln x_B + v_B)
\]
for banks A and B, respectively. These frontier values are indicated by the points marked with \( \otimes \) in the Figure 3.12. It is clear that the frontier output for Bank A lies above the deterministic part of the production frontier only because the noise effect is positive \((v_A > 0)\), while the frontier output for Bank B lies below the deterministic part of the frontier because the noise effect of negative \((v_B < 0)\). It can also be seen that the observed output of Bank A lies below the deterministic part of the frontier because the sum of the noise and inefficiency effects is negative (i.e., \( v_A - u_A < 0 \)).

Figure 3.12: Measuring technical efficiency using stochastic frontier analysis

The stochastic production frontier as specified in equation 3.26 has limited applicability in the banking efficiency analyses since it accommodate only single output, and thus, not consistent with multiple-outputs and multiple-inputs characteristics of the production process of banking firms. Researchers, therefore, generally use the stochastic cost frontier for obtaining the cost efficiency scores for individual banks. The stochastic cost frontier model can be written as:

\[
\ln C_i = c_i = C(y_i, w_i; \beta) + v_i + u_i = C(y_i, w_i; \beta) + \varepsilon_i
\]

(3.28)

where \( \ln C_i \) is the logarithm of observed cost of production for the \( i^{th} \) bank, \( C(\cdot) \) is the functional form of the core of the cost frontier, i.e., the deterministic part, \( y_i \)
is the logarithm of the output quantity, \( w_i \) is a vector of logarithm of input prices, \( \beta \) is a vector of unknown parameters to be estimated, \( u_i \) is the non-negative cost inefficiency effect and \( v_i \) is the random error which represents noise in the data and is usually specified as white noise. Additionally, \( u_i \) and \( v_i \) are considered as being independently distributed from each other. The minimum cost of a bank corresponds to the stochastic quantity \( \exp(\beta) + v_i \) and the measure of cost inefficiency is given by the ratio of minimal to actual cost, i.e.,

\[
CE_i = \frac{\exp(C(y_i, w_i; \beta) + v_i)}{\exp(C(y_i, w_i; \beta) + v_i + u_i)} = \exp(-u_i) \tag{3.29}
\]

Since the random variable \( u \) is non-negative, the value of the cost efficiency lies between 0 and 1, with the value of 1 representing totally cost efficient production.

In the stochastic frontier analysis, the estimation of cost efficiency relies on the choice of functional form and estimates of parameter \( \beta \). To estimate parameter(s) \( \beta \), one can make use of maximum likelihood (ML) method or, in some circumstances, corrected ordinary least squares (COLS) method. However, ML estimators have many desirable large sample (i.e., asymptotic) properties, and therefore, they are often preferred to other estimators as COLS (Coelli et al., 2005). The ML method aims to provide estimates of the production technology parameters (i.e., \( \beta \)) and bank-specific (in)efficiency.

In the ML estimation, introduced by Aigner et al. (1977) and Meeusen and van den Broeck (1977), one proceeds by specifying the likelihood function for the model (3.28) according to the assumption formulated about the distribution of the one-side disturbance \( u_i \). Indeed, while the noise component \( v_i \) is essentially always assumed to followed \( N(0, \sigma^2_v) \), there are several possibilities as regards the specification of the inefficiency term, which is usually assumed to follow a truncated-normal, half-normal, exponential or gamma distribution, anyhow holding the strong assumption that it is independent of the random deviation term and the other regressors. Meeusen and van den Broeck (1977) assume an exponential distribution for \( u_i \), and Aigner et al. (1977) discussed two distributions, the half-normal and the exponential, as possible candidates for the one-sided error term representing cost inefficiency. Stevenson (1980) first
suggested the truncated normal and Greene (1980, 1990, and 1993) has advocated the two-parameter gamma distribution. Lee (1983) suggests the four-parameter Pearson family of distributions. However, there is no a priori basis for choosing one distribution over another, and the worse thing is that different specifications have been found to give different estimates of cost inefficiency (Forsund et al., 1980; Schmidt, 1976). In practice, this choice is usually made for reasons of convenience and the most popular choice in the literature has been the half-normal distribution. In all these cases, it is possible to derive the distribution of the composed error term $\varepsilon_i(v_i + u_i)$ and go back from this up to the likelihood function of the log-linear model (3.28). The score functions allow then to derive the appropriate expressions for the numerical computation of $\beta$ estimator and the variance of composed error term, $\sigma^2$. 

Once the parameters are estimated, the centre of interest is the estimation of cost inefficiency, $u_i$. The $u_i$ must be observed indirectly since direct estimates of only $\varepsilon_i$ are available. The procedure for decomposing $\varepsilon_i$ into its two components $v_i$ and $u_i$ relies on considering the expected value of $u_i$ conditional upon $\varepsilon_i(=v_i+u_i)$. Jondrow et al. (1982) were the first to specify a half-normal distribution for the one-sided inefficiency component and to derive the conditional distribution $(u_i|v_i + u_i)$. Under this formulation of the half-normal distribution, the expected mean value of inefficiency, conditional upon the composite residual, is defined as:

$$E[u_i|\varepsilon_i] = \frac{\sigma \lambda}{1 + \lambda^2} \left[ \frac{\phi(\varepsilon_i \lambda / \sigma) - \varepsilon_i \lambda}{\Phi(-\varepsilon_i \lambda / \sigma)} \right]$$

(3.30)

where $\sigma^2 = \sigma_u^2 + \sigma_v^2$ captures inefficiency, for $\lambda = 0$ every observation would lie on the frontier (Greene, 1993). $\phi(.)$ and $\Phi(.)$ are, respectively, the probability density function and cumulative distribution function of the standard normal distribution.

The truncated-normal model is a more general form of the half-normal, where $u_i$ is distributed with a modal value of $\mu$ (Stevenson, 1980). The explicit form for the conditional expectation is obtained by replacing the $\varepsilon_i \lambda / \sigma$ in the
half-normal model with: \( u^*_i = \varepsilon_i \lambda / \sigma + \mu / \sigma \lambda \). If \( \mu \) is not significantly different from zero, the model collapses to the half-normal.

If an exponential distribution is imposed, with a density function of the general form \( f(u_i) = \theta \exp^{-\theta u_i} \), the conditional expectation is expressed as (Greene, 1993):

\[
E[u_i | \varepsilon_i] = \left( \varepsilon_i - \theta \sigma^2 \right) + \frac{\sigma \phi \left( \left( \varepsilon_i - \theta \sigma^2 \right) / \sigma_v \right)}{\Phi \left( \left( \varepsilon_i - \theta \sigma^2 \right) / \sigma_v \right)} \tag{3.31}
\]

in which \( \theta \) is the distribution parameter to be estimated.

The more general gamma distribution is formed by adding an additional parameter \( P \) to the exponential formulation, such that \( f(u_i) = \frac{\theta^P}{\Gamma(P)} u^{P-1} \exp^{-\theta u_i} \) with \( u_i \sim G[\theta, P] \) (Greene, 1990).

3.6.1 Panel data framework

Until the work of Pitt and Lee (1981), all efficiency measurement studies were cross-sectional. As pointed out by Schmidt and Sickles (1984), these models have three problems. First, the cost inefficiency of a particular bank can be estimated but not consistently. Second, the estimation of the model and the separation of inefficiency effect from statistical noise require specific assumptions about the distribution of inefficiency effect and statistical noise. Third, it may be incorrect to assume that inefficiency is independent of the regressors. They recommended that a rich panel data can overcome some of these difficulties, and listed out the following three principal benefits accruing to panel data in the context of production frontiers.

(i) No strong distributional specification is necessary for the inefficiency disturbance term,

(ii) When inefficiency is measured with panel data, it is estimated consistently as time \( T \to \infty \), and

(iii) Inefficiency can be measured without assuming that it is uncorrelated with the regressors.

The literature on SFA provides two types of panel data models, namely time-invariant and time-varying efficiency models.
3.6.1.1 Time-invariant efficiency models

The stochastic cost frontier with time-invariant efficiency can be written as:

\[ c_{it} = \alpha + x'_{it} \beta + v_{it} + u_i \]  

(3.32)

where \( c_{it} \) represents the logarithm of costs of the \( i^{th} \) bank at time \( t \) \((i=1,...,N; t=1,...,T)\), \( x_{it} \) are the regressors, \( \alpha \) and \( \beta \) are the coefficients to be estimated, \( v_{it} \) is the error term capturing noise. As the parameters of the model can be estimated in a number of ways, therefore, cost efficiency can be estimated in a different ways.

3.6.1.1.1 Fixed-effects model

The stochastic cost frontier with time-invariant efficiency as described above in equation (3.32) can be written as:

\[ c_{it} = \alpha_i + x'_{it} \beta + v_{it} \]  

(3.33)

In the equation (3.33), \( \alpha_i = \alpha + u_i \) is the common bank-effects of the fixed-effects model. The estimate of inefficiency \( \hat{u}_i \) is then defined as the distance from the bank-specific intercept \( \hat{\alpha}_i \) to the minimal intercept in the sample:

\[ \hat{u}_i = \hat{\alpha}_i - \min_{i} \hat{\alpha}_i \geq 0 \]  

(3.34)

Consequently, \( \hat{\alpha}_i = \min_{i} \hat{\alpha}_i \). The frontier is, thus, shifted to the bank with the smallest estimated intercept and any deviation from this intercept is interpreted as inefficiency. This model is either estimated by adding dummy variables for each of the banks and using OLS, or by performing the ‘Within transformation’ and applying OLS to the transformed model. There are some advantages in using the ‘Within estimator’ model. No assumptions need to be made about a specific distribution of the inefficiency term, as is the case with cross-sectional models. One need not assume that the inefficiency term is not correlated with the regressors. Further, the fixed-effects model has nice consistency properties.

However, the fixed-effects model has a potential serious drawback. Horrace and Schmidt (1996) found wide confidence intervals for the efficiency estimates based on the fixed-effects model. The estimation error and the uncertainty in the identification of the most efficient observation are among the explanations adopted to justify this result. A problem related to the ‘Within estimation’ is that if important time-invariant regressors are included in the frontier model, these will
show up as inefficiency in equation 3.33 (Cornwell and Schmidt, 1996). In other words, the fixed-effects ($\alpha_t$) capture both variations across banks in time-invariant cost efficiency and all phenomena that vary across banks but are time-invariant for each bank. Unfortunately, this occurs whether or not the other effects are included as regressors in the model. This problem can be solved by estimating model (3.32) in a random-effects context.

### 3.6.1.1.2 Random-effects model

In the random-effects model, the inefficiency terms $u_i$ are treated as one-sided i.i.d. random variables which are uncorrelated with the regressors $x_{it}$ and the statistical noise $v_{it}$ for all $t$. The random-effects model can be estimated using either least squares or maximum likelihood techniques. The least squares approach as proposed by Schmidt and Sickles (1984) involves writing the model in the form of the standard error-components model, then applying Generalized Least Squares (GLS). The ML approach involves making stronger distributional assumptions concerning the $u_i$'s. For example, Pitt and Lee (1981) assumed a half-normal distribution while Battese and Coelli (1988) considered the more general truncated-normal distribution. In the followings, we briefly discuss these approaches.

Schmidt and Sickles (1984) proposed a random-effects model to compute time-invariant efficiency. In the case of a cost function model, it is defined as:

$$c_{it} = \alpha + x'_{it}\beta + v_{it} + u_{it}$$

where $\mu = E(u_i) > 0$. In doing this transformation, $u^*_i = u_i - \mu$ has a zero mean by definition and usual GLS panel data techniques apply. Hence, the model can be rewritten as:

$$c_{it} = \alpha_i + x'_{it}\beta + v_{it}$$

where $\alpha_i = \alpha + u_i$. The estimate of inefficiency $\hat{u}_i$ is then defined (as in the fixed-effects model) as the distance from the bank-specific intercept to the minimal intercept in the sample:

$$\hat{u}_i = \hat{\alpha}_i - \min(\hat{\alpha}_i) \geq 0.$$
The frontier is thus shifted to the bank with the smallest estimated intercept. The estimate \( \hat{\alpha}_i \) is calculated by \( \hat{\alpha}_i = 1/T \sum_{t} \hat{\epsilon}_u \) for each \( i = 1, \ldots, N \), where
\[
\hat{\epsilon}_u = c_u - x_u' \hat{\beta}
\]
is the composed residual from regression.

Schmidt and Sickles (1984) point out that the random-effects model is more suitable for short panels in which correlation is empirically rejected. Hausman and Taylor (1981) developed a test, based on Hausman (1978), for the hypothesis that the error terms are uncorrelated with the regressors. If the null hypothesis of non-correlation is accepted, a random-effects model is chosen, otherwise a fixed-effects model is appropriate. The Hausman specification test is a test of the orthogonality assumption that characterizes the random-effects estimator, which is defined as the weighted average of the Between and Within estimator. The test statistic is
\[
H = \left( \hat{\beta}_{RE} - \hat{\beta}_{FE} \right) \left( \text{var}(\hat{\beta}_{FE}) - \text{var}(\hat{\beta}_{RE}) \right)^{-1} \left( \hat{\beta}_{FE} - \hat{\beta}_{RE} \right)
\]
(3.38) where \( \hat{\beta}_{RE} \) and \( \hat{\beta}_{FE} \) are the estimated parameter vectors from the random and fixed-effects models. Under the null hypothesis that the random effects estimator is appropriate, the test-statistic is distributed asymptotically as \( \chi^2 \) with degrees of freedom equal to the number of the regressors. Henceforth, large values of the \( H \) test statistic have to be interpreted as supporting the fixed-effects model. Hausman and Taylor (1981) developed a similar test of the hypothesis that the inefficiency terms are not correlated with the regressors.

The above panel data techniques avoid the necessity of distribution assumptions in both the specification and the estimation of stochastic frontier functions. However, if the latter are known, similar maximum likelihood techniques to the ones applied to the cross-sectional data can be applied to a stochastic production frontier panel data model in order to get more efficient estimates of the parameter vector and of the cost inefficiency scores for each productive unit. In this respect, Pitt and Lee (1981) derived the half-normal counterpart of Aigner et al.’s (1977) model for panel data, while Kumbhakar (1987) and Battese and Coelli (1988) extend Pitt and Lee’s analysis to the normal-truncated stochastic frontier panel data model. The inefficiency can be estimated by the conditional mean \( \hat{u}_i = E(u_i | \bar{u}_i) \), where \( u_i = \nu_i + \epsilon_i \) and \( \bar{u}_i = \frac{1}{T} \sum_{t=1}^{T} \hat{\epsilon}_u \).
proposed by Jondrow et al. (1982). Alternatively, Battese and Coelli’s (1988) estimator $E(\exp\{-u_{it}\}|\mathbf{x}_i^t)$ can be used to obtain point estimator bank-specific time-invariant cost inefficiency.

### 3.6.1.2 Time-variant efficiency models

The assumption of time-invariant efficiency seems to be very unreasonable with large panels, particularly if the operating environment is competitive. In such a situation, it is hard to accept the notion that inefficiency remains constant through many time periods (Kumbhakar and Lovell, 2000). Cornwell et al. (1990) and Kumbhakar (1990) were the first to propose a model with efficiency varying with time. The first study suggested the use of several estimation strategies, including fixed-effects and random-effects approaches, while the second study applied a maximum likelihood technique with the assumption that efficiency varies in the same way for all individuals.

If the assumption of a time-invariant inefficiency term is relaxed, the model to be examined is the following:

$$c_{it} = \alpha_{0t} + x_{it}' \beta + v_{it} + u_{it}$$

where $\alpha_{0t}$ is the cost frontier intercept common to all banks in period $t$ and $\alpha_{it} = \alpha_{0t} + u_{it}$ is the intercept for the $i^{th}$ bank in period $t$. Given that it is possible to estimate $\alpha_{it}$, the following estimates of the cost efficiency of each bank can be obtained:

$$\hat{u}_{it} = \alpha_{it} - \alpha_{0t}$$

where $\alpha_{it} = \min(\hat{\alpha}_{it})$. Since it is not possible to estimate all $\alpha_{it}$ because this would mean having to estimate additional $N \times T$ coefficients in addition to the parameter vector $\beta$, Cornwell et al. (1990) proposed the following functional form:

$$\alpha_{it} = \theta_{i1} + \theta_{i2} t + \theta_{i3} t^2$$

which reduces the problem to estimating $N \times 3$ parameters in addition to the ones contained in $\beta$. Substituting (3.41) in (3.39), we get the general model:

$$c_{it} = \theta_{i1} + \theta_{i2} t + \theta_{i3} t^2 + x_{it}' \beta + v_{it}$$

$$= w_{it}' \delta_i + x_{it}' \beta + v_{it}$$

132
When using a fixed-effects model, the estimation procedure starts by finding the ‘Within estimator’, \( \hat{\beta}_w \), then continues by applying OLS to a regression of the residuals \( y_{it} - \hat{x}_{it} \hat{\beta}_w \) to find estimates of the elements of \( \delta_i \) and then computing \( \hat{\delta}_i \) as \( w_i' \hat{\delta} \). Finally, estimates of inefficiency as in equation (3.40) will be obtained. Cornwell et al. (1990) consider the fixed-effects and random-effects approach. Since time-invariant regressors cannot be included in the fixed-effects model, they develop a GLS random-effects estimator for time-varying cost efficiency model. However, the GLS estimator is inconsistent when the cost inefficiencies are correlated with the regressors, therefore, the authors compute an efficient instrumental variables (EIV) estimator that is consistent in the case of correlation of the efficiency terms with the regressors, and that also allows for the inclusion of time-invariant regressors.

Lee and Schmidt (1993) propose an alternative formulation, in which the cost inefficiency effects for each productive unit at a different time period are defined by the product of individual cost inefficiency and time-effects,

\[
 u_{it} = \delta_i u_i
\]

where \( \delta_i \)'s are the time-effects represented by time dummies, and the \( u_i \) can be either fixed or random bank-specific effects.

On the other hand, if independence and distributional assumptions are available, ML techniques can also be applied to the estimation of stochastic frontier panel data models where cost inefficiency depends on time. Kumbhakar (1990) suggests a model in which the cost inefficiency effects assumed to have a half-normal distribution, vary systematically with time according to the following expression,

\[
 u_{it} = \delta(t) u_i , \text{ and } \delta(t) = [1 + \exp(bt + ct^2)]^{-1}.
\]

where \( b \) and \( c \) are unknown parameters to be estimated. In this model, the hypothesis of time-invariant cost efficiency can be verified by testing the hypothesis \( b = c = 0 \).
Another time-varying efficiency model was proposed by Battese and Coelli (1992) who assume cost inefficiency to be an exponential function of time and where only one additional parameter ($b$) has to be estimated,

$$u_{it} = \delta(t)u_i = \exp[-b(t-T)]u_i$$  \hspace{1cm} (3.43)

where $u_i$'s are assumed to be i.i.d. following a truncated-normal distribution. The drawback to these two models is that the inefficiency component follows a prescribed functional form, which might or might not be true. Especially in the case of the Battese and Coelli (1992) model, the evolution of the inefficiency component over time is monotonic, i.e., the inefficiency increases or decreases constantly over time, which need not hold in general. Cuesta (2000) specified a model of the form $u_{it} = \delta(t)u_i = \exp[-b_it(t-T)]u_i$. This model generalizes the Battese and Coelli (1992) model and allows the temporal pattern of inefficiency effects to vary across banks.

Kumbhakar and Hjalmarsson (1995) model the inefficiency term as:

$$u_{it} = a_i + \xi_{it}$$  \hspace{1cm} (3.44)

where $a_i$ is a bank-specific component which captures bank heterogeneity also due to omitted time-invariant variables, and $\xi_{it}$ is a bank time-specific component which has a half-normal distribution. The estimation of this model is in two steps. In the first step, either a fixed-effects model or a random-effects model is used to estimate all the parameters of the model $c_{it} = \alpha + x_i'\beta + v_{it} + u_{it}$, except those in equation (3.44). In the second step, distribution assumptions are imposed on $\xi_{it}$ and $v_{it}$. The fixed-effects ($\alpha + a_i$) and the parameters $\xi_{it}$ and $v_{it}$ are estimated by maximum likelihood, conditioned on the first step parameter estimates.

Battese and Coelli (1995) propose a model for stochastic cost inefficiency effects for panel data which includes explanatory variables. They modeled the inefficiency component as:

$$u_{it} = \delta'z_{it} + w_{it}$$  \hspace{1cm} (3.45)

where $u_{it}$ are cost inefficiency effects in the stochastic frontier model that are assumed to be independently but not identically distributed, $z_{it}$ is vector of variables which influence efficiencies, and $\delta$ is the vector of coefficients to be estimated. $w_{it}$ is a random variable distributed as a truncated-normal distribution.
with zero mean and variance $\sigma_u^2$. The requirement that $u \geq 0$ is ensured by truncating $w_u$ from below such that $w_u \geq -\delta' z_u$. Battese and Coelli (1995) underline that the assumptions on the error component $w_u$ are consistent with the assumption of the inefficiency terms being distributed as truncated-normal distribution $N^+(\delta' z_u, \sigma_u^2)$.

### 3.6.2 Stochastic distance functions

Traditional stochastic production frontier models are incapable to provide the technical efficiency scores when there is multiple-outputs since those models accommodate only single output. For computing technical efficiency of the banks in a multiple-outputs and multiple-inputs setting, the researchers are now widely using the stochastic distance functions (see, for example, Jiang et al., 2009; Cuesta and Orea, 2002; Rezitis, 2008; Koutsomanoli-Filippaki, 2009b). Stochastic distance function approach allows one to deal with multiple-inputs multiple-outputs (Coelli and Perelman, 1999) in the form of parametric distance functions, originally proposed by Shephard (1970). The basic idea is that in the case of a given production possibility frontier, for every bank the distance from the production frontier is a function of the vector of inputs used, $x$, and the level of outputs produced, $y$. For the output-oriented model, the distance function is defined as:

$$ D_o(x, y) = \min \{ \theta : y/\theta \in P(x) \} \quad (3.46) $$

where $D_o(x, y)$ is the distance function from the bank’s output set $P(x)$ to the production frontier. $D_o(x, y)$ is the non-decreasing, positively linearly homogeneous and convex in $y$, and decreasing in $x$ (Coelli and Perelman, 1999). $\theta$ is the scalar distance by which the output vector can be deflated (see Coelli, 2000) and can be interpreted as the level of efficiency. The output distance function aims at identifying the largest proportional increase in the observed output vector $y$ provided that the expanded vector $y/\theta$ is still an element of the original output set. If $y$ is located on the outer boundary of the production possibility set then $D_o(x, y) = \theta = 1$ and the utility is 100% efficient. On the other hand, values of $D_o(x, y) = \theta \leq 1$ indicate inefficient banks lying within the efficient frontier.
The input-oriented approach is defined on the input set \( L(y) \) and considers, by holding the output vector fixed, how much the input vector may be proportionally contracted. The input distance function is expressed by:

\[
D_i(x,y) = \max \left\{ \rho : x/\rho \in L(y) \right\}
\]

(3.47)

\( D_i(x,y) \) is non-decreasing, positively linearly homogeneous and concave in \( x \), and increasing in \( y \) (Coelli and Perelman, 1999). \( \rho \) is the scalar distance by which the input vector can be deflated. If \( D_i(x,y) = \rho = 1 \), then \( x \) is located on the inner boundary of the input set and the utility is 100% efficient.

In this approach, the first step is to determine the parametric relationship between inputs and outputs. The most common and appropriated functional form is the translog distance function. The translog input distance function \( D_i \) in its parametric form with \( M \) \((m = 1,\ldots,M)\) outputs and \( K \) \((k = 1,\ldots,K)\) inputs is specified as (Coelli, 2000):

\[
\ln D_i = \alpha_o + \sum_{m=1}^{M} \gamma_m \ln y_m + \sum_{k=1}^{K} \beta_k \ln x_k + \frac{1}{2} \sum_{m=1}^{M} \sum_{n=1}^{M} \gamma_{mn} \ln y_m \ln y_n + \frac{1}{2} \sum_{k=1}^{K} \sum_{l=1}^{K} \beta_{kl} \ln x_k \ln x_l + \sum_{k=1}^{K} \sum_{m=1}^{M} \delta_{km} \ln x_k \ln y_m
\]

(3.48)

In order to maintain the homogeneity and symmetry a number of restrictions need to be imposed. For homogeneity, the following restrictions have to be considered:

\[
\sum_{k=1}^{K} \beta_k = 1, \quad \sum_{j=1}^{K} \beta_{kj} = 0 \quad \text{and} \quad \sum_{k=1}^{K} \delta_{km} = 0
\]

For symmetry, two other restrictions have to be fulfilled:

\[
\gamma_{ma} = \gamma_{am} \quad \text{and} \quad \beta_{kl} = \beta_{lk}
\]

Imposing homogeneity restrictions by normalizing equation (3.48) by dividing the inputs by one of the inputs \( x_k \) delivers the estimating form of the input distance

\[
- \ln x_k = \alpha_o + \sum_{m=1}^{M} \gamma_m \ln y_m + \sum_{k=1}^{K-1} \beta_k \ln \left( \frac{x_k}{x_K} \right) + \frac{1}{2} \sum_{m=1}^{M} \sum_{n=1}^{M} \gamma_{mn} \ln y_m \ln y_n
\]

\[
+ \frac{1}{2} \sum_{k=1}^{K-1} \sum_{l=1}^{K-1} \beta_{kl} \ln \left( \frac{x_k}{x_K} \right) \ln \left( \frac{x_l}{x_K} \right) + \sum_{k=1}^{K-1} \sum_{m=1}^{M} \delta_{km} \ln \left( \frac{x_k}{x_K} \right) \ln y_m - \ln D_i
\]

(3.49)

Here, \( \ln D_i \) can be interpreted as error term which reflects the difference between the observed data realizations and the predicted points of the estimated function. Replacing \( \ln D_i \) by a composed error term (the stochastic error \( \nu_i \) and the
technical inefficiency $u_i$) yields the common SFA form. It can be estimated by a stochastic frontier production function defined as

$$y_i = f(x_i) + v_i + u_i.$$  

For $I(i=1,\ldots,I)$ banks, the econometric specification with $\ln D_i = v_i - u_i$, in its normalized form is expressed by:

$$-\ln x_{Ki} = \alpha_0 + \sum_{m=1}^{M} \gamma_m \ln y_{mi} + \sum_{k=1}^{K-1} \beta_k \ln \left( \frac{x_{ki}}{x_{Ki}} \right) + \frac{1}{2} \sum_{m=1}^{M} \sum_{n=1}^{M} \gamma_{mn} \ln y_{mi} \ln y_{ni}$$

$$+ \frac{1}{2} \sum_{k=1}^{K-1} \sum_{l=1}^{K-1} \beta_{kl} \ln \left( \frac{x_{kl}}{x_{ Ki}} \right) \ln \left( \frac{x_{ki}}{x_{Ki}} \right) + \sum_{k=1}^{K-1} \sum_{m=1}^{M} \delta_{km} \ln \left( \frac{x_{ki}}{x_{Ki}} \right) \ln y_{mi} + v_i - u_i \quad (3.50)$$

A distribution for $u_i$ has to be assumed in order to separate stochastic noise and inefficiency effects. One may assume, $u_i$ follows either the half-normal distribution 

$$\left( u_i \sim N^+\left[0, \sigma_u^2\right] \right)$$

or the truncated-normal distribution 

$$\left( u_i \sim N^+\left[0, \sigma_u^2\right] \right).$$

### 3.6.3 Marrying DEA with SFA

In recent years, attempts have been made by the researchers to marry DEA with SFA with the objective to reap the benefits of both the approaches in estimating banking efficiency. The first attempt in this direction has been made by Fried et al. (2002). The authors suggested three-stage approach to purging performance evaluation of environmental factors and statistical noise begins with DEA. In the second stage, SFA is applied to trace components of performance attributable to the operating environment of the unit, statistical noise, and managerial efficiency. In the third stage, data entered into DEA in Stage 1 are adjusted for the effect of the environment and statistical noise before repeating DEA. Thus, the evaluation emerging from the final stage DEA is said to represent managerial efficiency only.

In Stage 1, Fried et al. (2002) suggested the use of input-oriented variable returns-to-scale DEA with the conventional BCC model. The linear programming problem for envelopment form of BCC-I is given as:
\[
\min_{\theta, \lambda^i, \lambda^r} \theta - e \left( \sum_{i=1}^{n} s_i^- + \sum_{r=1}^{l} s_r^+ \right)
\]

subject to

\[
\sum_{j=1}^{n} \lambda_j x_{ij} + s_i^- = \theta x_{io}
\]

\[
\sum_{j=1}^{n} \lambda_j y_{rj} - s_r^+ = y_{ro}
\]

\[
\sum_{j=1}^{n} \lambda_j = 1
\]

\[
\lambda_j, s_i^-, s_r^+ \geq 0
\]

where \(x_{ij}\) is the amount of the \(i^{th}\) input used by the bank \(j\), \(y_{rj}\) is the amount of the \(r^{th}\) output produced by the bank \(j\), \(s_i^-\) and \(s_r^+\) represent input and output slacks, respectively. The optimal solution that emerges from the model (3.51) is the preliminary performance evaluation scores that are likely to be confounded by environmental effects and statistical noise.

In Stage 2, Fried et al. (2002) focus on radial slacks (i.e., input contraction) emerge from Stage 1 DEA rather than non-radial slacks (i.e., under-produced output). Using SFA, input slacks are regressed on observable environmental variables, and a composed error term that captures statistical noise due to measurement errors and managerial inefficiency. The main justification for SFA is an asymmetric error term that allows for identification of the one-sided error component (i.e., managerial inefficiency) and the symmetric error term component (i.e., statistical noise). The general function of the SFA regressions is represented in equation (3.52) for the case of input slacks:

\[
s_{i,j}^- = f^i \left( z_j; \beta^i \right) + v_{i,j} + u_{i,j}
\]

where \(s_{i,j}^-\) is the Stage 1 slack in the \(i^{th}\) input for the bank \(j\), \(z_j\) is the environmental variables, \(\beta^i\) is the parameter vectors for the feasible slack frontier and \(v_{i,j} + u_{i,j}\) is composed error structure where \(v_{i,j} \sim N(0, \sigma_v^2)\) represents statistical noise, and \(u_{i,j} \geq 0\) represents managerial inefficiency. Similarly, SFA regression for the case of output slacks can be given as follows:

\[
s_{r,j}^+ = f^r \left( z_j; \beta^r \right) + v_{r,j} + u_{r,j}
\]
The SFA regression model does not require specification of the direction of impact of environmental variables; this can be read from the signs of the parameters. Following each regression, parameters $\beta^i, \mu^i, \sigma^2_{v_i}, \sigma^2_{u_i}$ are estimated and permitted to vary across $N$ input slack regressions.

In Stage 3, the authors repeat the DEA of Stage 1 by replacing observed input data with input data that have been adjusted for the influence of environmental factors and statistical noise. Thus, the DEA analysis to emerge from Stage 3 represents performance due to managerial efficiency only.

The above three-stage analysis put forward by Fried et al. (2002) begins with traditional DEA using the BCC model. However, the BCC model while producing units-invariant (i.e., dimension free) radial inefficiency estimates, does not generate units-invariant estimates of non-radial inefficiency (or slacks). For consistent interpretation of DEA and SFA estimates, there is a need to choose a fully units-invariant DEA model. Such a solution exists within the slacks-based measure (SBM) of efficiency as suggested by Tone (2001) where it is possible to argue for either output maximization or input minimization. However, Fried et al. (2002) arbitrarily select input minimization, and thus, focus only on input slacks in Stage 2. Thus, Avkiran and Rowlands (2008) modified the research design put forward by the Fried et al. (2002), and proposed a more comprehensive analysis where total input and output slacks are measured simultaneously against the same reference set, facilitated by a non-oriented SBM model that is fully units-invariant.

The linear programming program for the non-oriented constant returns-to-scale SBM is shown below:

$$
\min \quad \rho = \frac{1 - \frac{1}{m} \sum_{i=1}^{m} s^-_i / x_{i0}}{1 + \frac{1}{s} \sum_{r=1}^{s} s^+_r / y_{r0}}
$$

subject to

$$
\sum_{j=1}^{n} \lambda_j x_{ij} + s^-_i = x_{i0} \quad \text{(3.54)}
$$

$$
\sum_{j=1}^{n} \lambda_j y_{ij} - s^+_r = y_{r0}
$$

$$
\lambda_j, s^-_i, s^+_r \geq 0
$$
where $s_i^-$ and $s_r^+$ represent input and output slacks, respectively. Alternatively, the model can be transformed into

$$
\rho = \left( \frac{1}{m} \sum_{i=1}^{m} \frac{x_{io} - s_i^-}{x_{io}} \right) \left( \frac{1}{s} \sum_{j=1}^{s} \frac{y_{ro} + s_r^+}{y_{ro}} \right)
$$

where the first term represents the mean contraction rate of inputs and the second term represents the mean expansion rate of outputs. In the model (3.54), the bank is deemed to be efficient if the optimal value for the objective function is unity. In other words, for a bank to be efficient, all optimal input slack (input excess) and output slack (output shortfall) must equal zero. In the alternative formulation represented by equation (3.55), SBM is the product of input and output inefficiencies. In this way, environmental variables are omitted in Stage 1 analysis.

In Stage 2, Fried et al. ignore output slacks and regressed only input slacks on the environmental variables because of their model’s input orientation. However, they do acknowledge that a case can be made where both input and output slacks are explained through SFA. As a modification, Avkiran and Rowlands (2008) focus on both input slacks and output slacks, and thus provide a more refined measure of organisational efficiency which can be incorporated into managerial decision-making with more confidence. Thus, Stage 2 analysis leads to an estimate of $m+s$ (i.e., inputs plus outputs) SFA regressions where slacks measured by SBM for each input (output) are regressed on environmental variables. Parameter estimates obtained from SFA regressions are used to predict input slacks attributable to the operating environment and to statistical noise. Thus, observed inputs can be adjusted for the impact of the environment and statistical noise by

$$
x_{i,j}^A = x_{i,j} + \left[ \max_j \left\{ z_j \hat{\beta}_i - z_j \hat{\beta}_i^b \right\} - \hat{\nu}_{i,j} \right] + \left[ \max_j \left\{ \hat{\nu}_{i,j} \right\} - \hat{\nu}_{i,j} \right]
$$

where $x_{i,j}^A$ is the adjusted quantity of $i^{th}$ input in $j^{th}$ bank, $x_{i,j}$ the observed quantity of $i^{th}$ input in $j^{th}$ bank, $z_j \hat{\beta}_i$ the $i^{th}$ input slack in $j^{th}$ bank attributable to environmental factors, and $\hat{\nu}_{i,j}$ the $i^{th}$ input slack in $j^{th}$ bank attributable to statistical noise.

Alternatively,

$$
x_{i,j}^A = \left( 1 + \text{AdjFactorEnvironment}_{s_i,j} + \text{AdjFactorNoise}_{s_i,j} \right) x_{i,j}
$$
where

\[
\text{AdjFactorEnvironment}_{x_{i,j}} = \left( \max_j \left\{ z_j \hat{\beta}^i \right\} \right) \left( 1 - \frac{z_j \hat{\beta}^i}{\max_j \left\{ z_j \hat{\beta}^i \right\}} \right), \quad \text{and}
\]

\[
\text{AdjFactorNoise}_{x_{i,j}} = \left( \max_j \left\{ \hat{\nu}_{i,j} \right\} \right) \left( 1 - \frac{\hat{\nu}_{i,j}}{\max_j \left\{ \hat{\nu}_{i,j} \right\}} \right)
\]

The first adjustment in the equation (3.56), \( \left[ \max_j \left\{ z_j \hat{\beta}^i \right\} - z_j \hat{\beta}^i \right] \), levels the playing field regarding the operating environment by placing all units into the least favourable environment observed in the sample. The second adjustment \( \left[ \max_j \left\{ \hat{\nu}_{i,j} \right\} - \hat{\nu}_{i,j} \right] \) places all units in the least fortunate situation (i.e., regarding measurement errors) found in the sample. Hence, banks enjoying relatively favourable operating environments and statistical noise would find their inputs adjusted upwards. Equation (3.57) is transformation of Fried et al.’s approach to adjusting inputs, where the researcher is better able to see the degree of adjustments attributable to the operating environment and statistical noise. This is achieved by taking ratios instead of differences and arriving at an adjustment factor which multiplies the observed input. The variable AdjFactorEnvironment represents the percent upward adjustment of the observed input for the impact of the environment, and the another variable, AdjFactorNoise, captures the percent upward adjustment attributed to statistical noise.

Similarly, observed outputs can be adjusted for the impact of the environment and statistical noise by

\[
y_{r,j}^A = y_{r,j} + \left[ z_j \hat{\beta}^r - \min_j \left\{ z_j \hat{\beta}^r \right\} \right] + \left[ \hat{\nu}_{r,j} - \min_j \left\{ \hat{\nu}_{r,j} \right\} \right]
\]

(3.58)

where \( y_{r,j}^A \) is the adjusted quantity of \( r^{th} \) output in \( j^{th} \) bank, \( y_{r,j} \) the observed quantity of \( r^{th} \) output in \( j^{th} \) bank, \( z_j \hat{\beta}^r \) the \( r^{th} \) output slack in \( j^{th} \) bank attributable to environmental factors, and \( \hat{\nu}_{r,j} \) the \( r^{th} \) output slack in \( j^{th} \) bank attributable to statistical noise.

Alternatively,

\[
y_{r,j}^A = \left( 1 + \text{AdjFactorEnvironment}_{y_{r,j}} + \text{AdjFactorNoise}_{y_{r,j}} \right) y_{r,j}
\]

(3.59)
where

\[
\text{AdjFactorEnvironment}_{r,j} = \left( \frac{z_j \hat{\beta}'}{y_{r,j}} \right) \left( 1 - \min_j \left\{ \frac{z_j \hat{\beta}'}{z_j \hat{\beta}'} \right\} \right),
\]

\[
\text{AdjFactorNoise}_{r,j} = \left( \frac{\hat{v}_{r,j}}{y_{r,j}} \right) \left( 1 - \min_j \left\{ \hat{v}_{r,j} \right\} \right)
\]

However, to use equations (3.56) or (3.58), it is necessary to distinguish input-sourced statistical noise \( (\nu_{i,j}) \) from managerial inefficiency \( (u_{i,j}) \) in the composed error term of the SFA regressions. Once \( \nu_{i,j} \) has been estimated for each unit, equations (3.56) or (3.58) can be implemented and observed input usage adjusted. The statistical noise attached to an input usage and in output generation is estimated residually by equations (3.60) and (3.61), respectively, as:

\[
\hat{E} \left[ \nu_{i,j} | \nu_{i,j} + u_{i,j} \right] = s_{i,j} - z_j \hat{\beta}' - \hat{E} \left[ u_{i,j} | \nu_{i,j} + u_{i,j} \right] \tag{3.60}
\]

and

\[
\hat{E} \left[ \nu_{r,j} | \nu_{r,j} + u_{r,j} \right] = s_{r,j} - z_j \hat{\beta}' - \hat{E} \left[ u_{r,j} | \nu_{r,j} + u_{r,j} \right] \tag{3.61}
\]

In Stage 3, SBM DEA analysis of managerial efficiency purged of the influence of operating environment and statistical noise. That is, in this final stage of the three-stage efficiency analysis, all units are re-evaluated after inputs and outputs have been adjusted for influences of operating environment and statistical noise.

3.7 Other parametric approaches

3.7.1 Distribution free approach (DFA)

The distribution free approach (DFA) was originally suggested in Schmidt and Sickles (1984) and introduced by Berger (1993). It assumes that efficiencies are stable over time while random error tends to average out, and thus, requiring little to be assumed about the distributional form of the efficiency measure and random error. In this sense, the methodology is relatively ‘distribution free’. DFA specifies a functional form for the efficiency frontier as does SFA, but it uses different way to separate the inefficiencies from the residual. In particular, it is based on a translog system of cost and input cost share equations, and it generates estimates of inefficiency for each bank in each time period. In DFA, the cost function can be estimated either by GLS, as in Schmidt and Sickles (1984) or by
using OLS, as in Berger (1993). Since DFA does not require any assumption about the distribution of either inefficiency or the random error, it is easier to implement than SFA. Some significant studies on the use of DFA in the banking industry are Berger (1993), Allen and Rai (1996), DeYoung (1997), Dietsch and Lozano-Vivas (2000), Rime and Stiroh (2003), Matousek and Taci (2004), and Weill (2007).

In order to estimate cost efficiency, let us suppose, there is a sample of banks indexed $i=1,\ldots, I$ in each of $T$ time periods indexed $t=1,\ldots, T$. For each bank, we observe expenditure $E_{it}$, a vector $y_{it}$ of outputs produced, and a vector $w_{it}$ of input prices paid. A translog system consisting of a cost equation and its associated input cost share equations can be written as

$$\ln E_{it} = \ln c(y_{it}, w_{it}, \beta') + v_{it} + u_{it} ,$$

$$\frac{w_{nit}x_{nit}}{E_{it}} = s_{nit}(y_{it}, w_{it}, \beta') + v_{nit}, \quad (n = 2, \ldots, N) \tag{3.62}$$

This system is estimated separately for each time period, and so the technology parameter vector has a time superscript. Within each time period the error vector $[v_{it}, v_{nit}]$ captures the effects of random statistical noise, and the error component $u_{it} \geq 0$ measures the cost of bank-specific cost inefficiency. Since $E(v_{nit}) = 0$, allocative efficiency is imposed, and so $u_{it}$ captures the cost of technical inefficiency only. Zellner’s Seemingly Unrelated Regression (SUR) estimator is used to estimate model (3.62). It is assumed that the $u_{it}$ are random effects distributed independently of the regressors. For each producer, the cost equation residuals $\hat{e}_{it} = \hat{v}_{it} + \hat{u}_{it}$ are averaged over time to obtain the average residual, $\hat{e}_{i} = (1/T) \sum_{t} \hat{e}_{it}$. On the assumption that the random-noise error component $v_{it}$ should tend to average zero over time, $\hat{e}_{i} = (1/T) \sum_{t} \hat{e}_{it} \approx \hat{u}_{i}$ provide an estimate of the cost inefficiency error component. To ensure that estimated cost inefficiency is non-negative, $\hat{e}_{it}$ is normalized on the smallest value, and we obtain

$$\ln \text{Eff}_{it} = \exp \left\{ -[\hat{e}_{i} - \min_{i}(\hat{e}_{i})] \right\}$$

where $\min_{i}(\hat{e}_{i})$ is the minimum value of the average error term for all banks in the sample. This estimator is similar to the GLS panel data estimator in which $u_{it}$ is
treated as a random-effects, and this similarity suggests that it is appropriated when \( I \) is large relative to \( T \) and when the \( u_i \) are orthogonal to the regressors. However, it differs from GLS in that the structure of the underlying production technology is allowed to vary through time. Berger (1993) noted that since the elements of \( v_{ii} \) may not fully cancel out through time for each producer, \( \hat{e}_i \) may contain elements of luck as well as inefficiency. To alleviate this problem, he recommended truncating the distribution of \( CE_i \) at its \( q \)th and \((1-q) \)th quantiles.

A disadvantage of DFA is the requirement that cost efficiency be time-invariant, and this becomes less tenable as \( T \) increases. However, DFA also has two distinct virtues. First, being based on a sequence of \( T \) separate cross-sectional regressions, it allows the structure of production technology to vary flexibly through time (although excessive variation in \( \hat{\beta}_i \) would be difficult to explain). Second, it does not impose a distributional assumption on the \( u_i \), it lets the data reveal the empirical distribution of the \( \hat{\epsilon}_i \equiv \hat{u}_i \).

3.7.2 Thick frontier analysis (TFA)

Berger and Humphrey (1991, 1992) developed another distribution free way of estimating cost frontiers using a single cross-section or a panel, the so-called ‘thick frontier analysis’ (TFA). The TFA specifies a functional form for the frontier cost function. In contrast to SFA which imposes arbitrary assumptions about the normal and half-normal distributions, and orthogonality between X-inefficiencies and regressors, TFA imposes no distributional assumptions on either inefficiency or random error. It assumes that deviations from predicted costs within the highest and lowest cost quartiles of observations (stratified by size class) represent random error, while deviations in predicted costs between the highest and lowest quartiles represent inefficiencies (Berger and Humphrey, 1997).

In TFA, instead of estimating a precise frontier bound, a cost function is estimated for the lowest average cost quartile of banks, which may be thought of as a ‘thick frontier’, where the banks exhibit an efficiency greater than the sample average. Similarly, the cost function for the highest average cost quartile is also estimated where the banks assumed to have less efficiency than the sample average. The difference in the predicted costs between these two ‘thick frontiers’ or ‘cost functions’ can be split into two factors. First is explained by market factors
related to the available exogenous variables and the second factor cannot be explained, i.e., the ‘inefficiency residual’. The predicted cost differences between highest and lowest cost quartiles between market factors and inefficiency residual can be decomposed as:

\[ \text{Diff} = \frac{A\hat{C}^{Q4} - A\hat{C}^{Q1}}{A\hat{C}^{Q1}} \]

where \( A\hat{C}^{Q1} \) and \( A\hat{C}^{Q4} \) represent the mean average cost of lowest and highest cost quartiles, respectively. The part of Diff owing to the exogenous market factors is given by:

\[ \text{Market} = \frac{A\hat{C}^{Q4*} - A\hat{C}^{Q1}}{A\hat{C}^{Q1}} \]

The remaining differences in average cost that cannot be attributed to the exogenous variables constitute the measured inefficiency residual as given by:

\[ \text{Ineff} = \frac{A\hat{C}^{Q4} - A\hat{C}^{Q1*}}{A\hat{C}^{Q1}} = \text{Diff} - \text{Market} \]

The distinct advantages of the thick frontier analysis is that (i) any number of exogenous variables may be added to the cost equation without changing the number of comparison banks or necessarily creating a downward bias in the inefficiency estimate, (ii) even if the errors terms within quartiles represent inefficiencies rather than random errors as maintained, the thick frontier approach remains a valid comparison of the average inefficiencies of high and low cost banks, and (iii) it reduces the effect of extreme points in the data, as can DFA when the extreme average residuals are truncated. However, an important caveat of TFA is that the measured efficiency under TFA is sensitive to the assumptions about which fluctuations are random and which represent efficiency differences (Berger, 1993). For example, if random errors follow a thick-tailed distribution and tend to be large in absolute value, while inefficiencies follow a thin-tailed distribution and tend to be small, then TFA may not yield precise estimates of the overall level of inefficiencies. However, Berger and Humphrey (1991) argued that ‘precise measurement is not our purpose rather, our goals are to get a basic idea of the likely magnitude of inefficiencies in banking and to identify their sources by decomposing them into several categories’.

### 3.7.3 Recursive thick frontier analysis (RTFA)

Recursive thick frontier analysis (RTFA) was developed by Wagenvoort and Schure (1999, 2006). It is panel estimation approach which avoids imposing
distributional assumptions on the inefficiency component of the error term. Unlike other frontier approaches, RTFA works well even if the number of time periods in the panel data set is small. RTFA is based on the assertion that if deviations from the frontier of X-efficient banks are completely random then one must observe for this group of banks that the probability of being located either above or below the frontier is equal to a half. This hypothesis can be tested for panel data sets but requires sorting of the full sample into a group of X-inefficient and X-efficient banks. The cost frontier is estimated using only the observations of the latter category by applying the ‘trimmed least squares’ (TLS) estimator (see Koenker and Bassett, 1978). Once the frontier is established, the X-efficiency can be computed as,

\[ XEFF_{it} = x_{it} \beta_{TLS}^{\ast} / y_{it} \]

Here, \( x_{it} \) represents a \( k \)-dimensional input bundle and \( y_{it} \) is a output bundle. \( \beta_{TLS}^{\ast} \) is TLS estimator of \( \beta \). RTFA allows X-inefficiency to vary over time, and be dependent on the explanatory variables of the frontier model.

Two important points are to be emphasized. First, in RTFA, the efficient banks are selected on the basis of their distance to the regression line instead of their average costs as in TFA. Second, even in the case where the observations of both the inefficient and efficient banks are drawn from a normal distribution, it is unlikely that the computed residuals of the regression equation are exactly normally distributed.

### 3.8 Comparison of DEA and SFA

In the banking efficiency literature, the most commonly used parametric approach is SFA and non-parametric approach is DEA. As already noted, both SFA and DEA have a range of advantages and disadvantages, which may influence the choice of methods in a particular application. The principal advantage of SFA is that it allows the test of hypothesis concerning the goodness of fit of the model. However, the major disadvantage is that it requires specification of a particular frontier function (like Cobb-Douglas or translog), which may be restrictive in most cases. Furthermore, the major advantage of the DEA is that it does not require the specification of a particular functional form for the technology. The main disadvantage is that it is not possible to estimate parameters for the model and hence impossible to test hypothesis concerning the performance of the model.
Table 3.18 provides a comparison of DEA and SFA on various aspects. It is significant to note here that though DEA and SFA are widely used by the researchers in their empirical analyses on the efficiency of banks, but no consensus has been reached in the literature about the appropriate and preferred estimation methodology (Iqbal and Molyneux, 2005; Staikouras et al., 2008).

<table>
<thead>
<tr>
<th>S. No.</th>
<th>Characteristic</th>
<th>Data Envelopment Analysis</th>
<th>Stochastic Frontier Analysis</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>Nature</td>
<td>Non-parametric approach</td>
<td>Parametric approach</td>
</tr>
<tr>
<td>2.</td>
<td>Functional specification</td>
<td>DEA does not require any <em>a priori</em> assumption about the selection of particular form of technical, cost or profit function relating to inputs and outputs.</td>
<td>SFA requires assumption about the particular form of technical, cost or profit function being estimated and the distribution of efficiency. Thus, the efficiency is then assessed in relation to this function with constant parameters and will be different depending on the chosen functional form.</td>
</tr>
<tr>
<td>3.</td>
<td>Frontier estimation</td>
<td>DEA constructs a piecewise linear-segmented efficient frontier with minimal assumption about the underlying technology making it less susceptible to specification error but with no scope for random error.</td>
<td>SFA constructs a smooth parametric frontier which accounts for stochastic error but requires specific assumptions about the technology and the inefficiency term which may be inappropriate or very restrictive (such as half-normal or constant inefficiency over time).</td>
</tr>
<tr>
<td>4.</td>
<td>Handling of inputs and outputs</td>
<td>DEA has the advantage that it is able to manage complex production environments with multiple inputs and output technologies.</td>
<td>SFA production frontier cannot manage multiple outputs. However, cost and profit frontier can accommodate multiple-outputs.</td>
</tr>
<tr>
<td>5.</td>
<td>Efficiency outcome</td>
<td>DEA efficiency estimates are based on a comparison of the input-output levels of an individual bank with those of a very small subset of efficient units.</td>
<td>SFA efficiencies are based on estimated average parameter values in the regression model.</td>
</tr>
<tr>
<td>6.</td>
<td>Random noise</td>
<td>DEA is a non-statistical (i.e., non-stochastic) methods and make no assumptions about the stochastic nature of the data. Thus, DEA is deterministic in nature. The non-stochastic nature of DEA implies that either the data are observed without error or the relationship between inputs and outputs is deterministic.</td>
<td>SFA is a statistical or econometric method tends to make assumptions about the stochastic nature of the data. Thus, SFA allows for statistical or random ‘noise’ in data.</td>
</tr>
<tr>
<td>7.</td>
<td>Sample size</td>
<td>DEA works particularly well with small samples.</td>
<td>SFA needs large samples to avoid lack of degrees of freedom.</td>
</tr>
</tbody>
</table>

*Source: Author’s elaboration*
3.9 Conclusions

The main objective of this chapter is to review the various analytical methods which are being used by the researchers to measure the efficiency of the banks. It has been observed that the frontier efficiency measurement methods made the traditional financial accounting ratios' analysis completely obsolete and outdated. From the deep inspection of literature, we note that data envelopment analysis (DEA) and stochastic frontier analysis (SFA) are the predominant frontier approaches that have attracted many empirical studies. Both the approaches have their distinct advantages and disadvantages. DEA has an advantage of computing efficiency scores in multiple-inputs and multiple-outputs production setting without specifying any functional form and distribution of the inefficiency term, but makes no room for noise and lacks hypothesis testing. On the other hand, though the efficiency estimates from SFA accommodate the noise, yet these estimates are very sensitive to the choice of functional form and distribution of the inefficiency term. In sum, no agreement has been reached on the superiority of one method over the others, and choice of a particular method is primarily guided by the data considerations and an individual preference.
Endnotes

1 Free disposability means that the destruction of goods is not expensive.
2 Convexity implies that the efficient frontier includes all linear combinations of dominant units.
3 DMUs are usually defined as entities responsible for turning input(s) into output(s), such as firms and production units. In the present study, DMUs refer to the individual banks. A DMU must have at least some degree of freedom in setting behavioural goals and choosing how to achieve them. In the present study, the term DMUs and banks are used interchangeably.
4 As with cost and revenue efficiency, the difference between SPE Type-I and Type-II is that Type-I model is traditional and commonly uses the original inputs/outputs values in constraints, while Type-II models use cost/revenue values of inputs/outputs in constraints.
5 Culture is the distinctive management and operational competencies that reflect the firm’s technology and processes.
6 Aside from the usual regularity axioms (i.e., ‘no free lunch’, the possibility of inactivity, boundedness, and closedness), FDH imposes only strong free disposability in inputs and in outputs, and so called free disposal (Wagenvoort and Schure, 1999). The former refers to the fact that any given level of output(s) remains feasible if any of the input is increased, whereas the latter means that with given input(s) it is always possible to reduce output(s).
7 DFA is also known as stability-over-time approach as it assumes that efficiencies are stable over time.
8 The frontier is called thick in order to indicate that best-practice banks are allowed to be positioned close to the frontier but not necessarily at the frontier.