CHAPTER–III
DATABASE AND METHODOLOGY

The research methodology and its implementation plan, research design, data issues and managing of these issues to proceed with the research plan are addressed in the ensuing description. The DEA models used in the study are also delineated.

3.1 STEPS OF RESEARCH METHODOLOGY

- Review of Literature
- Identification of Input and Output variables
- Data Collection Plan
- Sampling Plan
- Data Analysis Plan
- Data Issues
- Model Specification
- Software for Computation

Figure 3.1 presents the research methodology in a flow chart.

3.2 RESEARCH METHODOLOGY IMPLEMENTATION

3.2.1 Review of Literature

The studies of efficiency and productivity improvement have been examined to comprehend the drivers augmenting industry performance thereby leading to its growth and competitiveness. The appropriate strategies supporting such improvements were also spelt out.
Review of Literature

Identification of Input and Output Variables

Data Collection through Secondary Sources
1. Prowess Database
   (Data collected for the ten years 2001-02 to 2010-11 for 154 firms in the Indian IT Industry).
3. Annual Reports of NASSCOM (National Association of Software and Services Companies) and various issues of Strategic Review published by NASSCOM annually.

Data Analysis Techniques Used
1. Data Envelopment Analysis (DEA)
2. Multiple Regression

Data Issues
1. Identification
2. Management

Model Specification
BCC-I, CCR-O and BCC-O Envelopment DEA programs

Software for Computation
1. MAX DEA
2. SPSS (Statistical Package for Social Sciences)

Fig. 3.1: Flow Chart - Research Methodology
3.2.2 Identification of input and output variables

The diversity in the choice of financial variables made the task of selection of input and output variables difficult. Income statement data were used after an extensive and thorough literature review. The selection of the following variables is supported by empirical evidence.

Table 3:1 Table of Input and Output variables

<table>
<thead>
<tr>
<th>Variable</th>
<th>Explanation</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Inputs</strong></td>
<td></td>
</tr>
<tr>
<td>i) Compensations as a percentage to sales</td>
<td>Compensation paid to employees constitutes a major input in the IT industry. It is representative of labor employed. It is normalized by taking sales, and hence renders data across DMUs comparable. The difference in company size is hence taken care of.</td>
</tr>
<tr>
<td>ii) Depreciation</td>
<td>Expenditure on depreciation is deemed to represent capital usage/asset usage.</td>
</tr>
<tr>
<td>iii) Other expenses</td>
<td>It is the sum of all revenue expenses incurred by the company during the accounting period.</td>
</tr>
<tr>
<td><strong>Outputs</strong></td>
<td></td>
</tr>
<tr>
<td>i) Sales</td>
<td>It represents the revenue generated by the enterprise during the accounting period. The revenue is “measured by the charges made to customers or clients for goods supplied and services rendered to them”.</td>
</tr>
<tr>
<td>ii) Return on networth (RONW)</td>
<td>This output measures profitability. It is defined as: (100 \times \frac{\text{pat-nnrt}}{\text{average net worth}}) where pat is profit after tax; nnrt is net of non-recurring transactions.</td>
</tr>
</tbody>
</table>
3.2.3 Data Collection Plan

This study is based on secondary sources of data.

The Prowess Database published by CMIE is accessed to collect the financial data pertaining to the sample of firms selected in the Indian IT industry to workout technical efficiency of the industry over the decade (2001-02 to 2010-11).

The Global Information Technology Report prepared and published annually since 2001 by World Economic Forum, Geneva, in collaboration with INSEAD has been used. The Networked Readiness Index (NRI) furnishes a comprehensive assessment of the preparedness of the economies to reap benefits of technology for economic growth and productivity. The inclination for nations to leverage opportunities offered by ICT for development and competitiveness is gauged. It establishes an international framework for developing enabling factors. The role played by major stakeholders – individuals, businesses and governments in relation to three dimensions: the ICT environment, ICT usage and ICT readiness of stakeholders is studied.

![Diagram of Networked Readiness Index](image)

**Figure 3.2: The Framework of Networked Readiness Index**

Source: (Dutta & Mia, 2011)
The Annual Reports of NASSCOM (National Association of Software and Services Companies) and various issues of Strategic Review published annually by NASSCOM were used to further study the performance, growth and challenges faced by the Indian IT Industry.

### 3.2.4 Sampling Plan

A sample of 154 Indian IT Companies is selected to analyze the performance of Indian IT companies during the first decade of 21st century.

### 3.2.5 Data Analysis Plan

The technique of Data Envelopment Analysis (DEA) is used [see (Charnes & Cooper, 1985), (Banker et al., 1989) and (Seiford & Thrall, 1990)]. This technique is adept at estimating inefficiency in multiple input and output production correspondences. It relates the performance of each company in the industry to a piece-wise linear industry production frontier which is empirically estimated production function based on the inputs and outputs of the efficient companies. The benchmarks given by DEA model lay targets for improvements to be emulated by inefficient firms. The choice of DEA over conventional analytic tools is made because of its ability to capture a firm’s total financial situation. The financial data used served as proxies for the desired inputs and outputs analyzed for determining technical efficiency. Post DEA analysis examines the robustness of DEA results. The peer group analysis is carried out which tests the strength and vigour of DEA scores.

To study the efficiency of Indian IT industry at a macro level the Global Information Technology Report (GITR) 2010-11 is accessed. The Networked Readiness Index (NRI) for 138 nations prepared by incorporating 71 variables, published by World Economic Forum in partnership with INSEAD, is studied. The Networked Readiness Index is the aggregation of scores from the most disaggregated level of variables to the highest level of overall NRI scores. For each level arithmetic mean is used to aggregate the component of each category. The ranking of nations based on NRI score is examined. Moving a step ahead technical efficiency of nations is computed by DEA using environment and readiness indices as the two inputs and usage index as the output.
India’s reference set is obtained. To capture if technical efficiency, environment and readiness impacts usage across individuals, businesses and governments, multiple regression is computed. The significance of correlation between efficiency score and usage further enhances and captures the impact of ICT usage across countries.

3.2.6 Data Issues

The tool of DEA is vulnerable to data problems.

1. Data issues encountered

After collecting input/output data from the income statements of the selected sample of companies, some major difficulties were encountered. The assumption that DMUs selected for the study are homogeneous, if homogeneity is to mean that entities included in the evaluation set should have the same inputs and outputs in positive amounts is not met with. In our dataset collected from Prowess database, the problem of missing data entries were encountered. To aggravate the problem further, the output (RONW) column showed some negative entries violating the assumption of positivity of output values (i.e. all $y_{kj}$’s must be greater than equal to zero). At the very outset, the thought of excluding DMUs with blank and negative entries prevailed. An attempt to delete these from the input and output columns left us with sparse 19-20 DMUs with complete data. It hence left us to ponder, contemplate and seek alternatives. Intensive literature review paved the path to manage and resolve these data issues successfully.

2. Identification of the path to overcome data issues

Charnes, Cooper & Rhodes (1978) in an exemplar study necessitate that all input-output data are strictly positive. They left no room for blank entries. But, subsequent research detailed on the minimal data requirements and significantly relaxed the positivity condition. The minimalistic conditions are now known to be:

i. “At least one DMU consumes/produces every input and output.”

ii. “Each DMU consumes at least one input and produces at least one output” ((Fare & Grosskopf, 2004); (Shephard, 1970)).
Missing data are problematic in the applications of Data Envelopment Analysis. Normally, potentially important input and/or output variables have scant coverage and inadequate reporting of statistics pose a threat which needs to be adequately addressed. There are passing remarks in literature to this effect. Kau & Liu (2000) used fuzzy sets to model the ranges for missing data. The standard approach has been elimination of blank entries from the data matrices. By discarding DMU k or a relevant output j from the dataset, we shall lose valuable information of the production possibilities. If these blank entries do not result in any harm to the analysis, except for the missing information of the specific data entry, it is not advisable to eliminate available information. Kousmanen (2002) has shown the path to use DEA model without inflicting harm to the efficiency score, in the instance of incomplete data. He advocates the use of appropriate dummy values in place of blank data entries. When output j for the evaluated DMU k is missing, that is \( y_{kj} \) is a blank entry, it should be replaced with a zero. This shall have the same effect as omitting output j from the calculation of efficiency score for DMU k but keeping it available for remaining DMUs in the sample. By modeling missing data in this manner, one assumes that the DMU cannot weight the factors that are missing and hence \( u_j^* = 0 \). It is closely related to the treatment of zeroes in the data matrices (Thompson et al., 1993). The radial efficiency remains equal to that obtained from an assessment without the consideration of missing factors (Thanassoulis, Portela & Despic, 2007). Most DEA models accept this trick. A similar gimmick applies to missing inputs. By using a sufficiently large number, say \( M >> \max_{h,j} \{x_{hj}\} \), the problem of missing inputs in data matrices is appropriately addressed. To determine if M is large enough, a simple way is to check from the optimal solution if \( v_j^* = 0 \) for all \( x_{kj} = M \). Use of the approach put forth by Kousmanen (2002) is a bliss from the applied perspective. DMU specific modifications to the computation code (which can prove cumbersome) are automatically dispensed with.

3. Management of data issues

The negative values in the output (RONW) column were handled, thus by affecting data transformation. This property, referred to as translation invariance, is critical when the data has zero or negative values, which must be translated prior to
analysis with available software packages. A constant larger than the most negative value of RONW was added to the measured quantity of this output of all of the firms thus affecting a translation of the axis (in output space) so that the origin was shifted to a point in the positive orthant. Such transformation of the data is quite arbitrary, and is often carried out for computational convenience. This change of origin leaves the optimal solution of input-oriented BCC model unchanged. Hence BCC-I model is translation invariant with respect to outputs. This also helped us in the choice of the DEA model for the study. Many applications dealing with negative data which were frontrunners of the first theoretical papers on this issue are: (Pastor, 1996), (Zhu, 1996) [also (Seiford & Zhu, 2002)]; (Thore, Kozmetsky & Phillips, 1994) and (Lovell, 1995). It was in 1995 when the first paper devoted to negative values in DEA data was published. This question of invariance of DEA efficiency scores was addressed by Ali & Seiford (1990) who only considered the presence of zero values. Negative values were not mentioned in their paper. Pastor (1994) was the first to tackle this issue and provided translation invariance classification of 3 basic DEA models. He showed that an affine displacement does not affect the efficient frontier. His findings were published in (Lovell, Pastor & Turner, 1995) and (Pastor, 1996).

The missing entries in the columns for input (depreciation) and output (RONW) were handled thus by the use of the method proposed by (Kuosmanen, 2002) who was inspired by the problems that stemmed from empirical application of DEA to the problem at hand.

He proposed:

i. Insert a dummy value \( y_{kj} = 0 \) in the output matrix in place of a blank entry;

and

ii. Use some sufficiently large number, say \( M > \max_{h,j} \{ x_{hj} \} \) in place of a blank entry in the input matrix.

Almost all DEA models, both input and output oriented, accept this trick. It enables one to run DEA model automatically in such a way that blank entries do not count.
Hence, if DMUs are handicapped with missing data entries in input and output matrices, their elimination before DEA analysis is not the answer. These DMUs with missing data could be highly useful as reference or benchmark units, which span the efficient frontier, and may also represent desirable ‘diversity’, reflected in the wider spread of input-output mixes. But, at the same time it must not be forgotten that we assume the most pessimistic values for the missing entries, with zero outputs or arbitrarily large inputs, when true performance could have been better if reported. In turn, DMUs that openly report their data are fairly rewarded and also assessed in terms of a larger number of input-output dimensions. An important lesson, thus for DMUs is to pay more attention on reporting data in future! Whether the DMUs are actually encouraged and induced (by this treatment of missing data in efficiency assessment) to report truthful and complete data, can be probed by future research.

3.2.7 Model Specification

It has long been recognized that Data Envelopment Analysis (DEA) by its use of mathematical programming is particularly adept at estimating inefficiency in multiple input and output production correspondences. Following (Charnes, Cooper & Rhodes, 1978) and (Banker et al., 1989) a number of different DEA models have now appeared in literature (Seiford and Thrall, 1990; Seiford, 1996).

In the present study, to determine the drivers of technical efficiency of Indian IT industry, we have used the BCC-I envelopment DEA program, given by (Banker et al., 1989). This model explicitly assumes variable returns to scale (VRS).

The reasons for the choice of this model are:

i. Since data from financial statements were taken, there were some negative data entries in the output. The BCC-I model is translation invariant with respect to outputs, hence the choice of this model.

ii. The outputs, namely, return on net worth (RONW) and sales cannot be considered as proportionately dependent on inputs, namely, compensation as a percentage of sales, depreciation and other expenses. This calls for variable returns to scale.
Therefore, the choice and use of BCC model in the present study is reinforced.

Moreover, the input oriented model is chosen as the most appropriate for this application because only the inputs are controllable by the decision making units (DMUs).

The BCC-I envelopment DEA program helped us to compute technical efficiency of Indian IT firms.

Further, to examine the global competitiveness of Indian IT industry, BCC-O envelopment DEA program which assumes variable returns to scale (VRS) and CCR-O envelopment DEA program which assumes constant returns to scale (CRS) are used. Scale efficiency of nations is hence reported by computing the ratio of CRS efficiency to VRS efficiency. The output oriented models are used since the objective is to study maximization of individual, business and government ICT usage across nations.

**The BCC (Bunker, Charnes and Cooper) Model**

The input-oriented VRS envelopment DEA model (Banker, Charnes & Cooper, 1984) is expressed as follows:

\[
\min \theta_o - \varepsilon \left( \sum_{i=1}^{m} s_i^- + \sum_{r=1}^{s} s_r^+ \right)
\]

subject to

\[
\sum_{j=1}^{n} \lambda_{j} x_{ij} + s_i^- = \theta_o x_{io} \quad i = 1,2,\ldots, m;
\]

\[
\sum_{j=1}^{n} \lambda_{j} y_{ij} - s_i^+ = y_{ro} \quad i = 1,2,\ldots, s;
\]

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

\[
\lambda_{j}, s_i^-, s_r^+ \geq 0 \quad \forall i, r, j.
\]

where DMU_o represents one of the n DMUs under evaluation and x_{io} and y_{ro} are the i^{th} input and r^{th} output for DMU_o, respectively. The presence of the non-Archimedean \( \varepsilon \) in the objective function effectively allows the minimization over \( \theta_o \) to pre-empt the
optimization involving the slacks, $s_i^-$ and $s_r^+$. DMU$_o$ is efficient if and only if $\theta_o = 1$ and $s_i^- \neq 0$ and (or) $s_r^+ \neq 0$ for some $i$ and $r$.

The output oriented VRS envelopment DEA model (Banker, Charnes & Cooper, 1984) is expressed as follows:

\[
\text{max. } \phi_o + \varepsilon \left( \sum_{i=1}^{m} s_i^- + \sum_{r=1}^{s} s_r^+ \right)
\]

subject to

\[
\sum_{j=1}^{n} \lambda_j x_{ij} + s_i^- = x_{io} \quad i = 1,2,\ldots,m;
\]

\[
\sum_{j=1}^{n} \lambda_j y_{ij} - s_r^+ = \phi_o y_{ro} \quad r = 1,2,\ldots,s;
\]

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

\[
\lambda_j, s_i^-, s_r^+ \geq 0 \quad \forall i, r, j.
\]

where DMU$_o$ represents one of the $n$ DMUs under evaluation and $x_{io}$ and $y_{ro}$ are the $i^{th}$ input and $r^{th}$ output for DMU$_o$, respectively. The presence of the non-Archimedean $\varepsilon$ in the objective function effectively allows the maximization over $\phi_o$ to pre-empt the optimization involving the slacks, $s_i^-$ and $s_r^+$. DMU$_o$ is efficient if and only if $\phi_o = 1$ and $s_i^- \neq 0$ and (or) $s_r^+ \neq 0$ for some $i$ and $r$.

The CRR (Charnes, Cooper and Rhodes) Model

The output-oriented CRS envelopment model (Charnes, Cooper & Rhodes 1978) is expressed as follows:

\[
\text{max. } \phi_o + \varepsilon \left( \sum_{i=1}^{m} s_i^- + \sum_{r=1}^{s} s_r^+ \right)
\]

subject to

\[
\sum_{j=1}^{n} \lambda_j x_{ij} + s_i^- = x_{io} \quad i = 1,2,\ldots,m;
\]
\[
\sum_{j=1}^{n} \lambda_j y_{oj} - s_r^+ = \phi_o y_{ro} \quad r = 1, 2, ..., s;
\]

where DMU_0 represents one of the n DMUs under evaluation and x_{io} and y_{ro} are the i^{th} input and r^{th} output for DMU_0, respectively. The presence of the non-Archimedean \( \varepsilon \) in the objective function effectively allows the maximization over \( \phi_0 \) to pre-empt the optimization involving the slacks, \( s_i^- \) and \( s_r^+ \). DMU_0 is efficient if and only if \( \phi_0 = 1 \) and \( s_i^- \neq 0 \) and (or) \( s_r^+ \neq 0 \) for some i and r.

Post DEA, India’s reference set is examined. Multiple regression model is also used to study the impact of technical efficiency, environment and readiness of nations on ICT usage.

### 3.2.8 Data Processing Plan

The software MAX DEA was run to carry out DEA computations. Multiple Regression and Correlation were run on SPSS (Statistical Package for Social Sciences).

### 3.3 PROPOSED CHAPTER PLAN

I. Introduction

II. Review of Literature

III. Database and Methodology

IV. Structure and Performance of Indian IT industry

V. Industry competitiveness in Porterian terms

VI. Drivers of Technical Efficiency of Indian IT industry

VII. Competitiveness of Indian IT industry-A global perspective

VIII. Swot Analysis

IX. Summary and Conclusions
3.4 CONCLUSION

This chapter carefully presents the research methodology. The Data Issues encountered are identified and their management procedure laid down. The research design and the proposed chapter plan shall facilitate in meticulously proceeding and concluding the research work.