GRIPS Research Report Series I-2001-0003

Scale, Indivisibilities and Production Function in Data Envelopment Analysis

Kaoru Tone
&
Biresh K. Sahoo

National Graduate Institute for Policy Studies
2-2 Wakamatsu-cho, Shinjuku-ku,
Tokyo 162-8677, Japan

1 January 2002

GRIPS
NATIONAL GRADUATE INSTITUTE
FOR POLICY STUDIES

Scale, Indivisibilities and Production Function in Data Envelopment Analysis

Kaoru Tone

&

Biresh K. Sahoo

National Graduate Institute for Policy Studies
2-2 Wakamatsu-cho, Shinjuku-ku
Tokyo 162-8677, Japan

1 January 2002

Any comments are welcomed.

Scale, Indivisibilities and Production Function in Data Envelopment Analysis*

Kaoru Tone[a], *Professor*
&
Biresh K. Sahoo[b], *JSPS Postdoctoral Fellow*

Abstract

This paper critically re-examines the concept of returns to scale vis-à-vis economies of scale since the writings of Adam Smith by relating the former to the concept of *production unit* and the latter to the concept of *firm*. To date, the economic theory underlying DEA models has never been explored in any systematic manner. Hence one observes a significant divergence between the econometric and the DEA approaches for the estimation of production frontier. An attempt is made here to examine the DEA model of efficiency measurement and its application from an economic viewpoint. We show here that the presence of indivisibilities in all multi-stage production processes makes the technology structure non-convex, and therefore, the standard convex DEA production models (e.g., CCR and BCC) fail to exhibit scale economies due to such indivisibilities. However, the non-convex technology embedded in FDH model helps revealing process indivisibilities arising from task-specific processes whereas a homogeneous characterization of production function fails to achieve the same.

Key Words: Returns to scale; economies of scale; process indivisibility; data envelopment analysis.

---

* This research is fully supported by Grant-in-Aid for Scientific Research (C) Japan Society for the Promotion of Science.

[a] Corresponding author: National Graduate Institute for Policy Studies, 2-2 Wakamatsu-cho, Shinjuku-ku, Tokyo 162-8677, Japan. E-mail: tone@grips.ac.jp

[b] National Graduate Institute for Policy Studies, 2-2 Wakamatsu-cho, Shinjuku-ku, Tokyo 162-8677, Japan. E-mail: biresh@grips.ac.jp

Scale, Indivisibilities and Production Function in Data Envelopment Analysis

1. Introduction

The term 'economies of scale' is defined in the literature either in terms of physical output or in terms of cost of production. The neo-classical idea in terms of physical output is that a proportionate increase in the level of all inputs used in the process of production would result in a more than proportionate increase in the output. If the production is characterized by the notion of a neo-classical production function, then it is equivalent to saying that the production function is homogeneous of degree greater than one, which is otherwise called increasing returns to scale (IRS). Using the cost of production as the basis of defining scale amounts to saying that the unit cost of production decreases as the level of output expands, and is usually termed as economies of scale. If the cost of production is represented by a cost function derived from an underlying production function, then the two definitions are equivalent, and economies of scale would then represent cost savings due to IRS. However, the cost of production can also be a more general concept that includes savings in costs arising from sources like bulk buying at preferential lower prices, lower transport cost, lower advertising and other selling costs, none of which is directly related to the production process. Cost savings of this kind, if they exist, also reduce the overall average cost as output expands and should be recognized as scale effects. Thus, these two concepts measure scale economies arising from different sources.

The empirical estimation of scale, however, generally, uses either a total cost function (to test for declining average cost as an indication of scale) or a homogeneous production function (like Cobb-Douglas (C-D) or constant elasticity of substitution (CES)), whose degree of homogeneity indicates the presence or absence of scale effects. Either of the two approaches is generally taken to be a satisfactory way of empirical verification of scale. Whether they are taken to *highlight the same causal factors* is usually not mentioned. The first point made in this paper is that the failure to distinguish clearly between these two concepts of scale could lead to error in the interpretation of the results.

Economies of scale may arise on account of five sets of factors: (i) Returns to Scale (RTS); (ii) Behavior of overheads and indivisibility of factors of production; (iii) External and internal

economies; (iv) Nature of contracts between the `firm' and its constituents members/stakeholders as well as (v) their interrelationships that determine organizational efficiency. A detailed re-examination of the theoretical developments of these concepts in the following section shows that these two terms, Economies of Scale and Returns to Scale, have distinctive causative factors that do not permit them to be used interchangeably. In fact, we show in this paper that the tendency to use these two concepts as synonymous stems from narrowing down the very notion of a 'firm' to that of a `production unit' - an example of simplifying matters typical of neo-classical economics, whereas the modern firm (Aoki, 1990) is a complex phenomenon, a "Nexus of Contracts" which tries to economies on several counts, not mere the allocation of inputs.

The second concern of this paper is to address the question: *What light can either of the approaches mentioned above throw on the underlying sources of scale?* The answer is disappointing because of two fundamental problems: First the general nature of empirical research dealing with the estimation of cost/production function estimation is done at a level of aggregation that camouflages the sources of scale for particular industries. Very little insights can be inferred by observing some/ all encompassing measure of scale as to the nature of scale effects in that industry, thus making policy recommendations too general to be of practical use. The second more important problem is with the use of *homogeneous production function* to estimate RTS parameter. It is argued that such functional forms are far too narrow, perhaps even meaningless if the purpose at hand is to expose some of the well known arguments for increasing returns: *indivisibilities*. Very often, this is also a term that is used rather casually without going to the root what *kinds of indivisibilities* are actually operative at the production unit level. One of the greatest sources of confusion that emerges in relating *indivisibilities* and *scale* is again due to the very definition of scale adapted by neoclassical economic theory, which *necessitates constant factor proportions*. The requirement of equiproportionate changes in all inputs as a definition of scale is not here made because of empirical realities. Rather, it is argued that there are no compelling reasons for industries to maintain factor proportions constant during the process of expansion.

Finally, this paper aims at to point out one important kind of indivisibility that operates in most production process, but which cannot be captured by a homogeneous production function. This has to do with the production process and is called as "process indivisibility". This dimension to the indivisibility argument, although pointed out indirectly by economists like Marshall and Chamberlin, is shown to be *incompatible with the notion of a homogeneous production function.*

Since RTS is not defined for other non-homothetic functional forms except in a restrictive way, an alternative way of approaching the problem is suggested, which makes use of information on production as well as costs to describe scale effects in particular industries. We have shown here how the use of a nonparametric frontier estimated by data envelopment analysis[1] (DEA) could help revealing scale economies by capturing process indivisibilities arising from the multi-stage production, which a homogeneous production function might fail to do so.

The remaining part of the paper is unfold as follows: Section 2 deals with the historical evolution of the concept of economies of large scale production and the ideas associated with increasing returns. Section 3 develops a simple multi-stage model of a production process and shows how process indivisibilities arise and how they could lead to scale effects using DEA. Section 4 describes how scale effects occur in cement manufacturing based on empirical data on a representative sample of two mini-cement plants as an example of our arguments. Section 5 concludes.

2. Scale: a historical perspective

The Classicists defined the term `economies of scale' in the broadest sense. When the scale of operations is large, the cost advantages - due to division of labor (Adam Smith, 1791), effect of cooperation and team work (Karl Max, 1978), technological improvements (Marshall, 1920), technical and managerial improvements (Clark, 1923 and Robinson, 1935) - lead to a fall in the unit cost of production in the industry. Thus, the benefits of expansion, as expounded by these authors, flow from the many diverse components of what we label as a 'firm'. The emphasis here is not only on the technology but more on the entire gamut of organization, management, learning by doing, reorganization of inputs and other capabilities of the firm. This broader definition of scale is summed up by Silberston as follows: "economies of scale can be said to exist if an expansion in the volume of output produced results in a decrease in the unit cost of production when at each higher level of output, all possible adaptations in technology and organization have been carried through" (Silberston, 1972).

This broad definition of scale which is based on the concept of `firm' and which includes many dimensions other than production such as organization, financial capabilities etc. was lost in the

[1] The use of DEA to estimate production frontier, efficiency and returns to scale has later been extensively described in one subsection: Introduction to DEA models.

neoclassical formulation of scale. The concept of `firm' itself was never followed up and matters of equilibrium and markets became the preoccupation of the theorists. The 'firm' was increasingly treated as a technical unit, which converted a set of inputs into a single homogeneous output with little reference to its internal structure; and attention was diverted towards the study of perfectly competitive equilibrium and the theory of distribution.

Neoclassical definition: a critical review

It was Wicksteed (Stigler, 1946) who suggested that increasing returns should be attributed to proportionate increases rather than differential increases in all or some inputs. This was being done to facilitate the idea that efficient input proportions should change only if factor prices change with substitution between inputs. Thus, the requirement of proportionate changes in factors was not aimed at refining the concept of scale but to develop a 'theory of distribution' under a competitive market structure. However, Gold (1981) maintains that the shift of emphasis from `firm' to market structure indirectly led to the restatement of the concept of scale. The concern with the conditions of existence of competitive equilibrium, and the requirements of marginal productivity theory of distribution that returns to the factors of production in accordance with their marginal productivities would completely exhaust the total product, would coincide if the underlying production function were to be homogeneous. "Thus theoretical contributions had clarified the conditions under which scale increases could be integrated into, and even reinforce static economic theory. But the combination of restrictions involved in defining scale leaves unclear the sources of potential economies and even the relevance of this concept to any substantial sector of industry" (Gold, 1981, p.10).

Marshall disagreed with this new definition of equiproportionate changes by saying: "Increasing return is a relation between a quantity of effort and sacrifice on the one hand, and a quantity of output on the other. The quantities cannot be taken out exactly, because changing methods of production call for machinery, and for unskilled and skilled labor of new kinds and in new proportions" (Marshall, 1920, p.319).

The idea of a production function, first propounded by Wicksteed, became one of the most important tools of the neoclassical theory of production and distribution. The production function was quickly found a place in standard text books in microeconomic analysis. Specific functional forms that were employed in traditional literature were pioneered by Douglas (1948) and later on

by Arrow et al. (1961). The most common way in which RTS is characterized by production function formulations is through the class of homogeneous production functions. The degree of homogeneity of such a function being greater than one is representative of increasing returns to scale. Therefore, when scale is represented by a homogeneous production, it necessitates an equiproportionate increase in all inputs. Thus, Wicksteed's redefinition as alluded to earlier, has remained a dominant idea. However, Chamberlin (1948) was critical of this redefinition and said, "unless (entrepreneurs) harbor an interest in the mathematics of homogeneity, which submerges their ordinary entrepreneurial objectives, they will have no reasons to ..... maintain the proportions of factor constant" (p.143).

The most commonly used homogeneous production function to describe production technology in empirical literature is Cobb-Douglas production, which is given below:

$$Y = f(L,K) = AL^{\alpha}.K^{\beta},$$

where, A represents technology; L and K represent respectively labor and capital, and α and β represent respectively elasticity of output with respect to labor and capital. This production function is homogeneous of degree $(\alpha+\beta)$ because

$$f(\lambda L,\lambda K) = A(\lambda L)^{\alpha}.(\lambda K)^{\beta} = \lambda^{\alpha+\beta}.f(L,K); \lambda > 0.$$

It means that an equiproportionate change (10%, say) in all inputs leads output to change by $10^{\alpha+\beta}$. Here the degree of homogeneity, $\alpha+\beta$, represents RTS parameter. This production technology exhibits increasing returns to scale (IRS) if $\alpha+\beta > 1$, constant returns to scale (CRS) if $\alpha+\beta = 1$ and decreasing returns to scale (DRS) if $\alpha+\beta < 1$.

There are some economists such as Russell and Wilkinson (1979) who further narrow down the notion of production function to a consideration of only efficient technologies. Such technologies are defined so as to fulfil two conditions: not only should the technology dictate the *maximum* amount of output derived from given quantities of inputs, but the *minimum* quantities of inputs should be required to produce that amount of output. While the broad idea of a production function, being a "book of blue prints" of sorts, allows for the broader definition of scale to come through, the idea of scale benefits being reaped due to the adaptation of different methods available in the technology set could be represented by a homogeneous production function, only if, as a

coincidence, these different methods required equiproportionate changes in all inputs. Certainly, one cannot generalize, and this condition may be fulfilled, only as an exception than the rule.

Some authors[2] make a conceptual distinction between *returns to scale* and *returns to total outlay*. Returns to scale is defined with respect to equiproportionate changes in all inputs, but returns to total outlay need not imply that inputs increase equiproportionately; the increase in total outlay may be apportioned between inputs so as to lead to a differential increase in some or all inputs. This, in turn, suggests that expansion path of the firm need not be linear. Comparison is then made between returns to scale and returns to total outlay, the conclusion being that returns to total outlay would exceed returns to scale whenever the expansion path is non-linear. This comparison would be meaningless if returns to total outlay were to be the relevant way of measuring scale, and it is pointless comparing the non-linear expansion path with a hypothetical *scale-line*[3], which has no valid empirical support.

Returns to total outlay, while taking into account all possible sources of scale *within the production unit*, which is also the 'firm', cannot distinguish between various *sources of scale within the firm/industry*. However, *returns to total outlay and returns to scale coincide for homogeneous production functions* (which is shown in the next paragraph), and, therefore, such functional forms are usually assumed to adequately represent both economies of scale and returns to scale. A closer look at this argument reveals that this is a clear attempt to treat economies of scale synonymous with returns to scale. One can begin by observing that while many instances readily present themselves as contributing to scale, it is not clear at all how many instances would actually result in there being economies of scale for equiproportionate changes in all inputs.

To derive a cost function from a homogeneous production function (e.g., C-D production function) we solve the following decision problem in L and K:

$$Min_{L,K} \quad C = P_L L + P_K K$$

$$s.t. \quad Y = A L^{\alpha} K^{\beta}$$

$$L, K \geq 0.$$

[2] For instance, see Russell and Wilkinson (1979).

[3] The scale line refers to the line of expansion for equiproportionate increases in all inputs.

The solution to the above problem yields the following cost function[4]:

$$C = \bar{C}\, Y^{\frac{1}{\alpha+\beta}}, \text{where} \quad \bar{C} = \left(\frac{\alpha+\beta}{\alpha}\right)\left[\frac{1}{A}\left(\frac{\alpha}{\beta}\right)^{\beta} . P_L^{\alpha} . P_K^{\beta}\right]^{\frac{1}{\alpha+\beta}} = const.$$

We see here that as output expands, average cost (AC) falls, remain constant or increases if the production technology exhibits respectively IRS, CRS and DRS. This is because

$$\frac{dAC}{dY} = \left(\frac{1-(\alpha+\beta)}{\alpha+\beta}\right).\bar{C}\, Y^{\frac{1-2(\alpha+\beta)}{\alpha+\beta}} \lessgtr 0 \quad iff \quad (\alpha+\beta) \gtrless 1.$$

As regards the RTS possibilities this underlying cost structure contains essentially the same information as that in a homogeneous production technology if the former is derived from the latter. It means that the apparent distinction between returns to scale and economies of scale is absent, i.e., both these concepts are same in this representation of technology. This proves the fact that *returns to scale* and *returns to total outlay* coincide for all homogeneous structure of production technology.

Using the Shepherd's principle of duality, the cost function would exhibit declining long run average cost if the underlying production function did exhibit increasing returns. But scale effects are not confined to the production unit and can emerge from all other dimensions, which affect costs. These are obviously not being captured by the production function, and hence would not be reflected in the self-dual cost function. Therefore, if the cost function does indicate scale effects, then it would have to be from particular sources arising from the production unit and cannot be generally attributed, as is often the practice, to the various components of the 'firm' that contribute to scale. However, as we have just argued it is extremely unlikely that the actual production processes would result in production functions that are homogeneous.

Indivisibility argument to the explanation of scale

It remains to discuss the role played by the notion of *indivisibilies* as the principle way in which scale emerges. This concept has been used in the writings of Kaldor (1934), Joan Robinson (1969) and Chamberlin (1947-48). Although it has generated a lot of controversy in the nineteen forties, it continues to play an important role in the neo-classical explanation of scale. At the outset it ought to be mentioned that a review of the controversy is not attempted here, rather the resulting

[4] See the Appendix A for the derivation of the cost function.

understanding of the *kinds of indivisibilities* are the subject matter of attention. At a general level, indivisibilities often refer to the fact that certain capital equipments are available in certain capacities only, and if production is carried out at levels which are not at the designed optimum capacity levels, then the unit costs would be higher. This would also mean that there would be a fall in the unit costs if outputs were expanded. This is also referred to as overcoming the "lumpiness" problem.

How does the *indivisibility* argument fit in with the notion of fixed factor proportions in the neo-classical definition of scale? First, the long run average cost (LRAC) that is drawn as a smooth downward sloping curve rests on the envelope theorem. It is the "envelop" of the short-run average cost curves. For a continuous and smoothly declining LRAC, it is usually assumed that the "plant" possibilities are numerous. Plant does not refer to capital equipment but to the "aggregate of factors", also referred to as gross investment. In other words, the reference is to the capital embodied in capital equipment as well as the value of other factors of production. But the explanation of scale is by considering the "indivisibility" of the technique of production associated with a certain plant size, that is, the use of particular capital equipment is not equally efficient for smaller output levels. This, in turn, is attributed to indivisibility of technology that has been embodied in those particular equipments. Thus, the notion of homogeneous "capital" and homogeneous "labor" are indispensable to the arguments. The question remains whether this treatment of scale will be in conformation to equiproportionate changes in factors. Some of the illustrations of scale in Samuelson's Economics such as the automatic self adjusting mechanisms, or Allport's example of the introduction of "new" factors such as computers (Allport and Stewert, 1978) clearly violate the requirements of fixed factor proportions as they are labor displacing.

Another form of *indivisibility* by which scale may emerge is to consider the use of equipment, which has the characteristics of incorporating proportionately less "capital" than its contribution to capacity when output is expanded. Physical capital equipments in the form of cylinders, pipes, vessels, etc., would all exhibit the well known engineers' 0.6 rule of thumb, i.e., a 100% increase in capacity leads to only 60% increase in costs. This, along with proportionate increase in all other raw materials and labor, would lead scale effects. Such effects would be purely due to the physical properties of materials and should be treated as natural sources of scale. Even here there are difficulties: while each individual piece of capital equipment may exhibit such properties, it does not follow that when used in specific combinations with other factors of production, the aggregate

of "capital" would show equiproportionate increases along with other factors of production for it to be representable by a homogeneous production function. In fact, there is the question of whether these advantages would be so pervasive so as to lead to scale effects at all. Gold (1981) observes that "such a relationship may hold, of course, in respect to some kinds of facilities, especially in respect to the construction of hollow shells such as tanks, furnaces, boilers, pipes, and some simple buildings. But fundamental shortcomings narrowly restrict the range of its applicability to complex production equipment ... *(because of)* the tendency for larger units to require intricate arrays of interconnected, precisely designed functioning components, as well as costly instrumentation, controls, and ancillary facilities" (p.12) (italic words are ours).

To conclude, *indivisibilities* have been used to provide a rationalization of the greater productive efficiency of large-scale operations in a framework that leaves much to be desired. What seems to be more important is to pin down the specific ways in which increased efficiency could be achieved and the potential for reorganization of inputs, which can emerge due to indivisibility of specific inputs.

Empirical evidence on scale

Haldi and Whitcomb (1967) regress various capital equipments on their respective output capacities to see whether costs go less than proportionately than capacity. They found that in basic industries, such as petroleum refining, primary metals and electric power, economies of scale are found up to very large plant sizes. The resulting saving in initial capital investment cost are important sources of scale economies. Also, significant economies are found to be present in operating expenses for labor, supervision and maintenance. The best-known study using the engineering approach to identify the sources of scale is by Chenery (1949). In this approach each element of the production process has been studied to discover the relation between inputs and outputs at different scales for that process.

In the study on scale by Soni and Jani (1987), the data clearly show that the ratio of capital to labor is changing over years, which suggests that the underlying production function is not homogeneous. Even if it is admitted that most industrial processes result in non-homogeneous production functions, the problem remains in defining scale for such functional forms. The translog production function is one such functional form for which the scale parameter varies across input proportions as well as with volume of output. However, the value of scale parameter is again

calculated for *given proportions* of inputs and varies whenever these input proportions change. For each input combination, the scale parameter is defined *as if that input proportion were to remain constant.* If returns to scale were estimated using the translog production function, any single measure of scale would require parametric restrictions on the functional form to bring it down to the conventional definition. The alternative is to divide the range of expansion of firms' output whenever the factor ratios change and estimate scale parameter using the appropriate functional form for each such division. The production function at each range of output would give different estimates of scale parameters. Here the empirical production frontier is piecewise linear and the definition of RTS through equiproportionate change in all inputs still holds true on each particular facet of the production surface. It is, therefore, suggested to preferably use DEA pioneered by Charnes et al. (1978) and Banker et al. (1984) to estimate RTS for all such facets so as to make us enable to draw inference regarding the RTS possibilities of the industry. This approach may be considered superior to estimate scale parameter than to have the one estimated by the homogeneous production function over the entire range of output.

Economic rationale of proportionality postulate

Let us illustrate Koopmans' (1957) [who is one of the founders of the formal approach to production theory] proportionality postulate with an example. If an activity A employing two inputs, one unit of labor, L and one unit of capital, K yields one unit of output, Y, then the activity λA using $\lambda.L$ labor and $\lambda.K$ capital produces $\lambda.Y$ units of output. In terms of our neoclassical production function formulation, if $1.Y = f(1.L, 1.K)$, then $\lambda.Y = f(\lambda.L, \lambda.K)$, which is the case of CRS representation. However, this existence of CRS is argued against here through an empirical proposition. The example here is that if one man ($1.L$) and one shovel ($1.K$) yield 1 acre of ploughed land ($1.Y$), then 10,000 men ($10{,}000.L$) with 10,000 shovels ($10{,}000.K$) yield 10,000 acres of ploughed land ($10{,}000.Y$). This augmented activity, 10,000A is, however, quite possible, but not efficient, and hence, is not a point on the production frontier (why?). By 10,000.K we do not mean 10,000 shovels but one tractor if the amount of capital embodied in this tractor is that of 10,000 shovels. Similarly, a few skilled labors can be substituted for 10,000 men. This new input combination will possibly yield more than 10,000 acres of ploughed land, which is the case of IRS. In any production process, we think of augmenting capital in two ways, one by mere replication and the other by reconfiguration. Replication means increases of original capital good by integer multiples (creating units identical to those already in use) whereas reconfiguration of capital means altering its physical specifications. In response to the need to alter the rate of output, a differently

constituted machine (tractor in our example) may be used. However, in case of mere replication, the production function will always exhibit CRS.

To Koopmans, this proportionality postulate implies that IRS is impossible unless there is an *indivisibility* or lumpiness in one or more of the inputs. If there were no such *indivisibility* of inputs, the technique that proves superior at the higher scale could, always be subdivided proportionately to produce efficiently at the lower scale. This argument reveals that *indivisibilities* are common in real world and CRS is not typical of most real-world production functions.

If these observations are put together, we are led to the fact that any meaningful notion of returns to scale in production is to do with the fact that there is some kind of *indivisibility* in the activities associated with the *production process*, and that there is also a 'hierarchy of techniques' available to produce different scales of output, both of which could lead to scale effects, although any one of them existing without other would lead to scale. But these facts do not depend upon any notion of a production function, much less a homogeneous production function to understand and measure scale. It would be worth observing that these ideas take us back to the broader definition of scale discussed by the classicists.

With regard to the theoretical implications of these ideas, it is clear that what is being suggested is that the scale-line of the firm is nonlinear. In accordance with the views expressed by Robinson (1969), it appears to be the only view that is consistent with empirical facts. Nonlinear scale-line and the reasons for such expansions have not been given adequate treatment in the literature, mainly because of the preoccupation with homogeneous functional form; it is as if mathematical convenience dictated which direction theory would take. In empirical work one needs to pin the non-linearity of the scale-line to the specific notion of indivisible activities within the productive process and the adaptation of different techniques.

A useful way of doing this is adopting a different way of looking at production which views the production as a *task-specific process* in which production is broken into its various principle stages. The idea is to bring out the inherent 'hidden' *indivisibilities* of the activities associated with the production process by observing the task-length associated with each stage. The main observation is that production process usually consists of more than one stage of production, and the task-lengths associated with various stages need not be equal. This is because different pieces of capital equipment used at different stages of production processes serve different purpose and are designed

with respect to that purpose at hand with the existing technical know-how. This simple observation seems to be enough to generate a nonlinear scale-line.

Indivisibility and technology set

The parametric technologies that are embedded in C-D/CES production functions maintain the implicit maintained hypothesis of convex structure, which, in turn, requires that inputs are not indivisible. Baring a few[5], most of the nonparametric technologies is also based on the convexity assumption. The only difference between these two types of technologies is that the former is differentiable with all its arguments while the latter is not. These technologies require convexity assumption in order to ensure the diminishing marginal rate of technical substitution, leading to *isoquant* convex to the origin. The assumption of convexity holds under two conditions[6]: 1) when output produced is very large, and 2) when small production processes can be replicated. However, most of the real-life production processes fail to satisfy these stringent criteria on account of several reasons, of which *indivisibility* is one of the obvious candidates. We then need to redefine the production technology in the presence of *indivisibility* in our production process.

For the same of simplicity, let us assume that there are only two techniques available to produce a particular output, y_0. A decision making unit (DMU_1) employs technique A that uses x_{11} amount of input type 1 and x_{21} amount of input type 2 to produce output, y_0. Similarly, DMU_2 employs technique B that uses x_{12} amount of input type 1 and x_{22} amount of input type 2 to produce the same output. These two techniques are represented as follows:

$$A = \begin{bmatrix} x_{11} \\ x_{21} \\ y_0 \end{bmatrix} \qquad B = \begin{bmatrix} x_{12} \\ x_{22} \\ y_0 \end{bmatrix}$$

Here, techniques A and B are efficient in the sense that the inputs they employ are minimum, and outputs they produce are the maximum. It is to be mentioned here A and B cannot be scaled down to produce less but can be scaled up by integer values to produce more. Employing the assumption of *convexity* and *free disposability*, the structure of technology can be represented as follows:

$$T_{convex}=\{(x_1,x_2,y_0): \lambda x_{11}+(1-\lambda)x_{12}\le x_1, \lambda x_{21}+(1-\lambda)x_{22}\le x_2, \lambda y_0+(1-\lambda)y_0 \ge y_0, 0\le\lambda\le 1\}.$$

---

[5] See, for example, among others, Tulken and Vanden Eeckaut (1995) and Kerstens and Vanden Eeckaut (1999).

[6] See Varian (1992) for its proof.

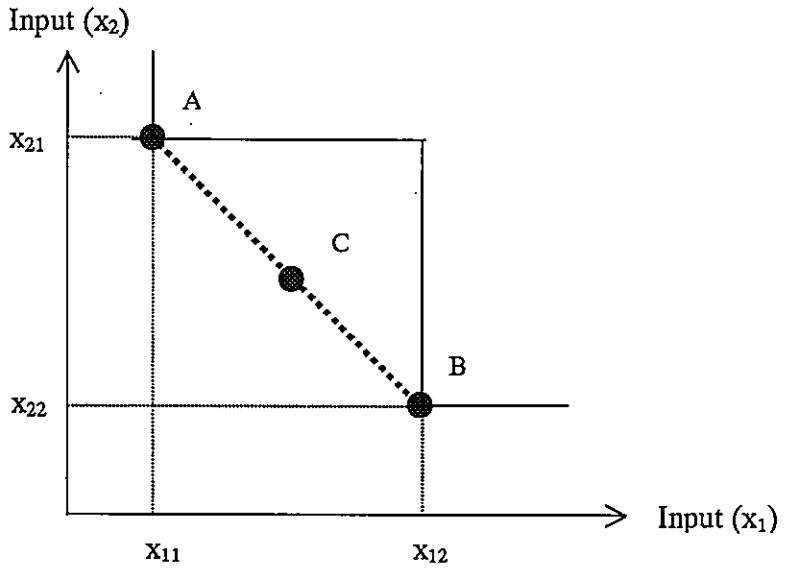


Figure 1: Convex and non-convex isoquants

For $\lambda = ½$, this technology implies that there exist another technique (C) that can use half of A and half of B to produce y_0 [i.e., $C = (½)A + (1-½).B$][7]. However, this is not possible because techniques, A and B cannot be scaled down because of their indivisibility nature. Either one can use technique A or technique B to produce y_0 ($\lambda = 0$ or 1). So in the presence of *indivisibility*, the structure of technology becomes non-convex in nature and can be modified as

$$T_{non\text{-}convex}=\{(x_1,x_2,y_0):\lambda x_{11}+(1-\lambda)x_{12}\le x_1,\lambda x_{21}+(1-\lambda)x_{22}\le x_2,\lambda y_0+(1-\lambda)y_0\ge y_0,\lambda\in\{0,1\}\}.$$

So, on generalization where there are n DMUs with each DMU_j, $j = 1,2,\ldots,n$ produces s different outputs, y_{rj} ($r = 1,2,\ldots,s$) using m different inputs, x_{ij} ($i = 1,2,\ldots,m$), $T^{8}_{\text{non-convex}}$ becomes

$$T_{non\text{-}convex}=\left\{(x,y):\sum_{j=1}^{n}\lambda_j x_{ij}\le x,\sum_{j=1}^{n}\lambda_j y_{rj}\ge y,\sum_{j=1}^{n}\lambda_j=1,\lambda_j\in\{0,1\}\right\}.$$

[7] Any combination of these two techniques (e.g., C in our example) could also mean that one can use A half time and B half time to produce y_0. However, it is not clear here whether both A and B will be in operation simultaneously. In any case, this interpretation does not make any sense in real-life situations.

[8] In DEA literature this is referred to as free disposal hull (*FDH*) technology (T_{FDH}) constructed on the basis of 'dominance relation' between observed input-output vectors. See Tulken and Vanden Eeckaut (1995) for its detailed discussion. However, ours is the first attempt in this paper to link the notion of *indivisibility* with the structure of technology, i.e., non-convexity.

Introduction to DEA Models

Before we move to the next section let us briefly present here the various DEA models that are required for the estimation of returns to scale and efficiency. Retaining the original symbols for input, output and DMUs, the input-oriented[9] CCR model (Charnes et al., 1978), based on the assumption of CRS is as follows:

$$\min \ \theta$$

$$s.t. \ \sum_{j=1}^{n} \lambda_j x_{ij} \leq \theta x_{i0}, \quad i = 1,2,.....m;$$

$$\sum_{j=1}^{n} \lambda_j y_{rj} \geq y_{r0}, \quad r = 1,2,.....s;$$

$$\lambda_j \geq 0, \quad j = 1,2,.....n.$$

where x_{i0} and y_{r0} are the i^{th} input and r^{th} output respectively for DMU_0 under evaluation. A score of unity for θ indicates that DMU_0 is efficient and any score, which is less than unity, turns DMU_0 as inefficient. However, if $\sum\lambda_j^* = 1$ in any alternate optima, then CRS prevails on DMU0; if $\sum\lambda_j^* < 1$ for all alternate optima, then IRS prevails; and if $\sum\lambda_j^* > 1$ for all alternate optima, then DRS prevails on DMU_0.

The input-oriented BCC model (Banker et al., 1984), based on the assumption of variable returns to scale (VRS), is obtained by adding convexity constraint to the original CCR model as follows:

$$\min \ \phi$$

$$s.t. \ \sum_{j=1}^{n} \lambda_j x_{ij} \leq \phi x_{i0}, \quad i = 1,2,.....,m;$$

$$\sum_{j=1}^{n} \lambda_j y_{rj} \geq y_{r0}, \quad r = 1,2,.....,s;$$

$$\sum_{j=1}^{n} \lambda_j = 1;$$

$$\lambda_j \geq 0, \quad j = 1,2,.....,n.$$

The dual of the above input-oriented BCC model is

[9] Though the determination of returns to scale is conditional on the choice of a measurement orientation, we feel to maintain input orientation in all the DEA models in our study for one simple reason: Real world managers are never given a bundle of inputs and told to produce the maximum output from it. Instead they are given output targets and told to produce it most efficiently, i.e., with minimum inputs. See also Sengupta (1987, p.2290) who argues in favor of input orientation in all practical situations.

$$\max \sum_{r=1}^{s} u_r y_{r0} + u_0$$

$$st. \quad \sum_{r=1}^{s} u_r y_{rj} - \sum_{i=1}^{m} v_i x_{ij} + u_0 \le 0, \quad j = 1,2,.....n;$$

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

$$u_r, \ v_i \ge 0, \quad and \ u_0 : free$$

In BCC model the sign of u_0^* (which is the optimal value of u_0) determines the nature of RTS for DMU_0. If $u_0^* = 0$ in any alternate optimal then CRS prevails on DMU_0, if $u_0^* > 0$ in all alternate optimal then IRS prevails and if $u_0^* < 0$ in all alternate optimal then DRS prevails on DMU_0. Banker et al. (1984) show that the above definition is consistent with the neoclassical characterization of RTS by Panzar and Willig (1977).

Färe et al (1985) introduced the following 'scale efficiency index' method, based on non-increasing returns to scale (NIRS), to determine the nature of RTS as follows:

$$\min f$$

$$st. \quad \sum_{j=1}^{n} \lambda_j x_{ij} \le f x_{i0}, \quad i = 1,2,.....m;$$

$$\sum_{j=1}^{n} \lambda_j y_{rj} \ge y_{r0}, \quad r = 1,2,.....s;$$

$$\sum_{j=1}^{n} \lambda_j \le 1;$$

$$\lambda_j \ge 0, \quad j = 1,2,.....n.$$

If $\theta^* = \phi^*$ iff DMU_0 exhibits CRS; otherwise if $\theta^* < \phi^*$ then DMU_0 exhibits IRS iff $\phi^* > f^*$ and DMU_0 exhibits DRS iff $\phi^* = f^*$.

These three different RTS methods are equivalent to estimate RTS parameter (Banker et al. (1996b) and Färe and Grosskopf (1994)). In empirical applications one, however, finds that the CCR and BCC RTS methods may fail when DEA models have alternate optima. However, the scale efficiency index method does not suffer from the above problem and hence is found robust.

In the light of all possible multiple optima problem in the CCR and BCC methods, Banker and Thrall (1992) generalized by introducing new variables u_0^+ and u_0^- which represent optimal solutions obtained by solving the dual of the input-oriented BCC model with one more constraint $\sum u_r y_{r0} + u_0 = 1$ and replacing the objective function in this model by either $u_0^+ = \max u_0$ or $u_0^- =$

max – u_0. They show here that IRS operates iff $u_0^+ \geq u_0^- > 0$, DRS operates iff $0 > u_0^+ \geq u_0^-$ and CRS operates iff $u_0^+ \geq 0 \geq u_0^-$.

Banker et al. (1996b) point out that the concept of RTS is unambiguous only at point on the efficient facets of production technology. So the RTS for the inefficient units may depend upon whether the efficiency estimation is made through an input-oriented or output-oriented model. A detailed method of doing so is found in the studies of Banker et al. (1996a) and Tone (1996).

Most of these studies, however, have unnoticed one important point that RTS is a local measure, though Banker et al. (1989, p.145) in their study argue that RTS in the BCC model hold only in DMU's current position. This point was clearly observed in the study of Golany and Yu (1997) who offers at least a partial remedy to the identification problems associated with the local nature of the RTS property by performing a sensitivity analysis based on observing the RTS behavior of the DMU being analyzed in two specified directions (one to the right (GY1) and another to the left (GY2)) as follows:

GY1:

$$\min \beta_0 - \varepsilon\left[\sum_{i=1}^{m} s_{i0}^- + \sum_{r=1}^{s} s_{r0}^+\right]$$

$$st. \quad \sum_{j=1}^{n} \lambda_j x_{ij} + s_{i0}^- = \beta_0 x_{i0}, \quad i = 1,2,.....m;$$

$$\sum_{j=1}^{n} \lambda_j y_{rj} - s_{r0}^+ = (1+\delta) y_{r0}, \quad r = 1,2,.....s;$$

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

$$\lambda_j \geq 0, s_{i0}^-, s_{r0}^+ \geq 0, \quad j = 1,2,.....n.$$

GY2:

$$\max \alpha_0 + \varepsilon\left[\sum_{i=1}^{m} s_{i0}^- + \sum_{r=1}^{s} s_{r0}^+\right]$$

$$st. \quad \sum_{j=1}^{n} \lambda_j x_{ij} + s_{i0}^- = (1-\delta) x_{i0}, \quad i = 1,2,.....m;$$

$$\sum_{j=1}^{n} \lambda_j y_{rj} - s_{r0}^+ = \alpha_0 y_{r0}, \quad r = 1,2,.....s;$$

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

$$\lambda_j \geq 0, s_{i0}^-, s_{r0}^+ \geq 0, \quad j = 1,2,.....n.$$

Here, δ is assumed to be a very small arbitrary positive number. Based on solutions of GY1 and GY2, the following procedures for estimating RTS are suggested:

Step 1. Solve GY1 to determine the RTS to the right of DMU_0:
Step 1(i). $1+\delta > \beta_0^* > 1 \Rightarrow$ IRS.
Step 1(ii). $1 \geq \beta_0^* \Rightarrow$ DMU_0 is inefficient.
Step 1(iii). $1+\delta = \beta_0^* \Rightarrow$ CRS.
Step 1(iv). $1+\delta < \beta_0^* \Rightarrow$ DRS.
Step 1(v). No feasible solution $\Rightarrow$ there is no data to determine the RTS to the right of DMU_0.

Step 2. Solve GY2 to determine the RTS to the left of DMU_0:
Step 2(i). $1 > \alpha_0^* > 1-\delta \Rightarrow$ DRS.
Step 2(ii). $\alpha_0^* \geq 1 \Rightarrow DMU_0$ is inefficient.
Step 2(iii). $1-\delta = \alpha_0^* \Rightarrow$ CRS.
Step 2(iv). $\alpha_0^* < 1-\delta \Rightarrow$ IRS.
Step 2(v). No feasible solution $\Rightarrow$ there is no data to determine the RTS to the left of DMU_0.

Most of these DEA models suffer from the occurrence of frequently observed zero optimal weights to some inputs/outputs, which might be very important to a particular DMU. In order to avoid this situation, Tone (2001) suggests a new variant of DEA model [Weight Restriction (WR) Model] to estimate RTS under weight restrictions as follows:

$$
\begin{aligned}
&\max z = uy_0 - u_0 \\
&s.t. \quad vx_o = 1 \\
&\quad -vX + uY - u_0 e \leq 0 \\
&\quad vP \leq 0 \\
&\quad uQ \leq 0 \\
&\quad v \geq 0, u \geq 0,
\end{aligned}
$$

where the matrices P and Q are associated with weight restrictions. Usually, he observes that the optimal solution in his model is degenerate and u_0^* is not unique. Then, analogous to the Banker and Thrall Procedure (1992), he suggested the following theorem for a WR-efficient DMU as follows:

i) if u_0^+ (sup of u_0^*) < 0, then RTS is increasing,
ii) if u_0^- (inf of u_0^*) < 0 < u_0^+ (sup of u_0^*), or $u_0^- = u_0^+ = 0$, then the RTS is constant,
iii) if 0 < u_0^-, then the RTS is decreasing.

The problem with CCR and BCC model is that they assign unity efficiency score even when the DMUs are not efficient in Koopmans sense. To deal with this problem, Tone (2001) and Cooper et al. (1999) suggested the following slacks-based measure (SBM) of efficiency:

$$
\begin{aligned}
&\min \tau = t - \frac{1}{m}\sum_{i=1}^{m} S_i^- / x_{i0} \\
&s.t. \quad 1 = t + \frac{1}{s}\sum_{r=1}^{s} S_r^+ / y_{r0}, \\
&\quad tx_0 = X\Lambda + S^-, \\
&\quad ty_0 = Y\Lambda - S^+, \\
&\quad \Lambda \geq 0, \quad S^- \geq 0, \quad S^+ \geq 0, \quad t > 0.
\end{aligned}
$$

This model declares a DMU efficient only when it is efficient in Koopmans sense, i.e., slacks inputs and outputs are zero. It is to be noted here that the optimal SBM ι^* is not greater than the optimal CCR θ^*, i.e., $\iota^* \leq \theta^*$.

All of these DEA models discussed so far are based on the maintained hypothesis that technology set is convex. However, as we have shown earlier, in the presence of *indivisibilities* the technology set is no longer convex. These convex models then fail to enable us to correctly determine the RTS possibilities along the frontier. Kerstens and Vanden Eeckaut (1999) proposed a more general method by considering variations on the existing FDH technology that is suitable for all reference technologies. Introducing the assumption of CRS into the FDH model yields the following input-oriented mix-integer non-linear programming problem[10] (Kerstens and Vanden Eeckaut, 1999):

FDH-CRS:

$$\min \quad \rho$$

$$s.t. \quad \sum_{j=1}^{n} z_j x_{ij} \leq \rho x_{i0}, \quad i = 1,2,.....,m;$$

$$\sum_{j=1}^{n} z_j y_{rj} \geq y_{r0}, \quad r = 1,2,.....,s;$$

$$\sum_{j=1}^{n} \lambda_j = 1;$$

$$\lambda_j \in \{0,1\},$$

$$z_j = \delta.\lambda_j,$$

$$\delta \geq 0, \quad \delta : \quad free.$$

Here, λ is the only activity operating subject to a non-convexity constraint and one re-scaled activity, z allowing for any scaling of the observations spanning the frontier. Similarly, introducing non-increasing returns to scale (NIRS) and non-decreasing returns to scale (NDRS) into FDH technology yield the same model with a change in restriction on σ, i.e., $0 \leq \sigma \leq 1$ and $\sigma \geq 1$ respectively in FDH-NIRS and FDH-NDRS models. So a DMU is efficient and considered lying on the boundary of the technology set if $\rho = 1$ and it is *undominated*. By *undominated* it is meant that there does not exist any other DMU with less of any input with same output, or any DMU with more of any output with same input, or any DMU with less input and more output. Letting the optimal values of DMU_0 in FDH-CRS, FDH-NIRS and FDH-NDRS models be respectively $\rho_{FDH\text{-}CRS}$, $\rho_{FDH\text{-}NIRS}$ and $\rho_{FDH\text{-}NDRS}$, its RTS is characterized locally by:

[10] This model turns to simple FDH model when the restriction, $z_j = \sigma.\lambda_j$ is removed and $z_j = \lambda_j$.

$$CRS \Leftrightarrow \rho_{FDH-CRS} = \max \{\rho_{FDH-CRS}, \rho_{FDH-NIRS}, \rho_{FDH-NDRS}\};$$
$$IRS \Leftrightarrow \rho_{FDH-NDRS} = \max \{\rho_{FDH-CRS}, \rho_{FDH-NIRS}, \rho_{FDH-NDRS}\}; or$$
$$DRS \Leftrightarrow \rho_{FDH-NIRS} = \max \{\rho_{FDH-CRS}, \rho_{FDH-NIRS}, \rho_{FDH-NDRS}\}.$$

Now let us turn to our proposed multi-stage production model.

3. The model

Assume that in a simple economy there is only one production process available, which consists of four different stages to get final output, say, Y. This method uses two factors of production, machines and labor to get the final output. Let M_1, M_2, M_3 and M_4 denote per unit of machines of type 1, 2, 3, and 4 respectively used in the first, second, third and fourth stages of production. The functioning of all the four machines used in various stages is as follows:

Stage 1: Some raw materials are processed by machine, M_1 with fixed number of labor. The resulting intermediate output is then entered into Stage 2. This stage takes 1 machine-hour of M_1 to process one unit raw material worth of output. If the unit time taken is one workday (12 hours), then Stage 1 can convert 12 units raw material worth of output.

Stage 2: The intermediate product so produced at Stage 1 is then enters into Stage 2 where it is further processed with machine, M_2 with fixed number of labor, and it takes 2 machine-hours of M_2 per unit raw material worth of output. Thus, the capacity of the machine M_2 is 6 units raw material worth of output per workday.

Stage 3: The job of Stage 3 is similar to that of Stage 2. Since the capacity of M_3 3 machine-hours per unit output, this stage produces 4 units raw material worth of output per workday.

Stage 4: Finally, the intermediate product produced in Stage 3 comes to Sage 4 where it is again processed with machine M_4 with fixed number of labor. This stage takes 1 machine-hour per unit of final output Y to be produced so that the total capacity is 12 units per workday.

If the production of final output has to be carried out in strict sequence, then during the whole work-day, 12 units raw materials worth of output will come out from Stage 1 and 4 units of final output would emerge with 6 units raw material worth of output awaiting at Stage 2 and 2 units raw material worth of output awaiting at Stage 3. It is worth noting here that if the total final output is 4 units per work-day, then Stage 1 has only 4 hours of work, Stage 2 has 8 hours, Stage 3 has full

12 hours of work, and Stage 4 has 4 hours of work with idle capacities of 8, 4, 0, and 8 hours respectively in existing at stages 1, 2, 3, and 4. If all the stages have to be fully utilized, then there must be two groups of Stage 2 types tasks and three groups of Stage 3 types tasks arranged to work simultaneously to be able to produce 12 units worth of final output. So the facilities of all these stages have to be fully utilized if the organization of sorts discussed above is carried through. This is shown in the Figure 2.

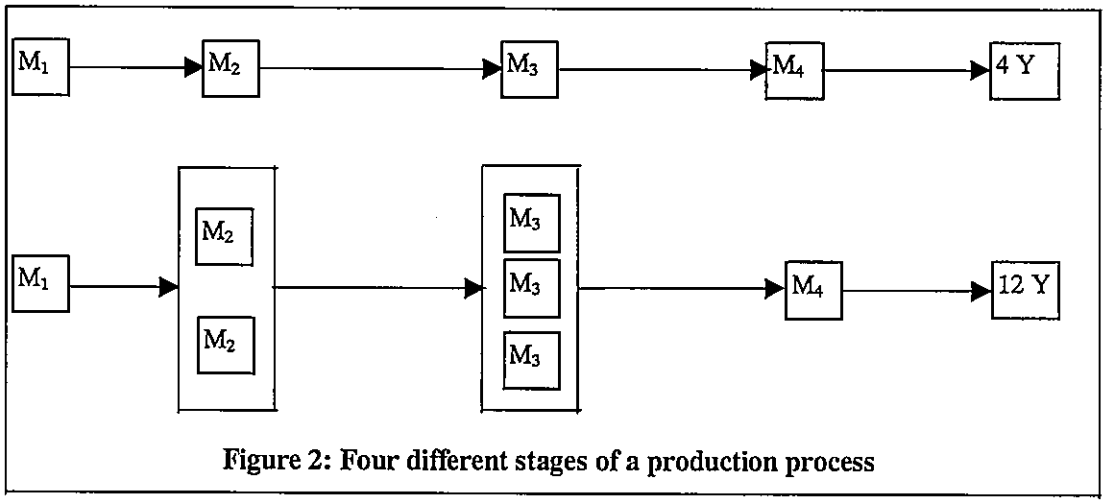


Figure 2: Four different stages of a production process

It is important to note that there are two alternative ways to produce 12 units of final output. One way is that simply replicating three times the process used to produce 4 units would result in 12 units of final output. The other way is to organize production as discussed above to take the maximum advantage of the existing idle capacities to obtain 12 units of final output. In the former case, the corresponding total cost will be three times that of the process used to produce 4 units whereas in the later case the total cost would be less than the three times of the original cost. What we infer from the organization of production is that the tripling of output does not necessitate the tripling of all the inputs. It follows that the total cost increases in less than proportion to total output, which is an indication of scale.

Given the process of production described above, if the technology is to be represented by *production function* it would have to be in terms of machine and labor. We here abstract labor from the production function because the ratio of labor to machine does not vary in this representation of technology. It is shown in Table 1 how the production is expanded by taking the maximum advantage of idle capacities existed in various stages of production.

Table 1: Organization of a particular production process

Stage 1 M_1	Stage 2 M_2	Stage 3 M_3	Stage 4 M_4	Final Output (Y)
1(8)	1(4)	1(0)	1(8)	4
1(6)	1(0)	2(6)	1(6)	6
1(4)	2(8)	2(0)	1(4)	8
1(0)	2(0)	3(0)	1(0)	12
2(0)	4(0)	6(0)	2(0)	24

Note: The figures in the brackets indicate the number of idle hours per workday.

It is clear from the Table 1 that the maximum amount of output obtained from the utilization of machines, M_1, M_2, M_3, and M_4 is 4 units. Here there are idle capacities of 8 hours in stage 1, 4 hours in stage 2, and 0 hour in stage 3 and 8 hours in the final stage. In order to produce more than 4 units but less than and/or equal to 6 units, one more machine M_3 in Stage 3 is needed, i.e., one M_1, one M_2, and two M_3 and one M_4, are needed to produce 6 units of output. Again even here, idle capacity exists in all the stages except in stage 2. So in order to produce more, i.e., up to 8 units, one more M_2 will be added along with factor combination (M_1, M_2, $2M_3$, M_4) used for the production of 6 units of output. At this stage of production, idle capacity exists in all the stages excepting at stage 3. If further production, i.e., up to 12 units is desired, then one more machine M_3 is needed in stage 3 along with the input combination (M_1, $2M_2$, $2M_3$, M_4) used for the production of 8 units. Now to obtain any level of output greater than 12 units requires proportionate increase in all inputs because there exists no idle capacity in any of these stages. For example, replicating the process used to produce 12 units two times would result in the production of 24 units of output with no idle capacities available in any of the stages.

Clearly, we observe that these five production possibilities constitute the vertices of the production frontier, because they are efficient in Koopmans sense. And the structure of this frontier is non-homogeneous (e.g., like that of a FDH frontier). Looking at the movement of the input and output vectors along the frontier, we expect that the *production function* exhibits increasing returns to scale up to the range of 12 units of output. And as regards the *expansion path*, it is nonlinear because doubling of output does not necessitate the doubling of all inputs during this output range even though doubling of all inputs leads to doubling of output. However, the latter production possibilities are not efficient and hence do not operate on the efficient frontier. So, this observation calls into question the ability of the homogeneous production function to capture scale effects if scale arises in this fashion.

One problem with this formulation of production technology in terms of physical inputs (machines) may arise when the physical inputs change over the scale of production with the premise that higher capacity machines are found to be efficient in the greater range of production[11]. There may arise question as to what does the 'doubling of inputs' mean when higher range of output requires a change in technique in the production process, i.e., higher capacity machines are used when the scale of production is large. This problem might be solved if we take `capacity' instead of machine as factor input. And it is to be noted here that the capacity (in terms of machine-hours) available, but not capacity spent, should be taken into consideration, otherwise the production function would always exhibit CRS.

Estimation of production function and returns to scale

It remains to be seen that if the notion of production function captures all such scale effects as propounded by classicists, then the relevant question is how to reveal this frontier from the *technology set* when this technology embodies efficient as well as inefficient production possibilities. We find in the literature that there are a number of approaches to frontier estimation used to evaluate productive performance, and DEA is one of them. Each of these approaches is consistent with the definition of production/cost/revenue/profit function as a boundary function. Based on the observed best practice, DEA, in principle, proves to be consistent with the more demanding standards set by boundary production function.

In the econometric literature, one generally imposes a functional form for the underlying production technology and then tests whether the industry as a whole exhibits CRS. However, as we have discussed earlier, the empirical technology comprises a number of piecewise-linear segments, indicating RTS to vary at each successive linear segment. It is worth noting here that the econometric approach to the estimation of scale fails to make one realize how RTS change along the boundary. On the contrary, DEA employs the postulate of minimum extrapolation from the observed data to estimate the production frontier. Unlike the econometric technique, it floats different piecewise-linear surfaces in different segments of the production technology, indicating that the DMUs operating on the efficient facets of production technology are characterized by varying returns to scale. In this connection we find the studies of Banker et al. (1986) and Sahoo et

[11] Capacity varies more than proportionately with cost due to engineers' 0.6 rule of thumb. Otherwise, mere replication of smaller production process would prove more economical.

al. (1999) interesting. The former study contrasts DEA and translog estimates of hospital production correspondence, and its translog results suggest CRS while its DEA indicates the presence of both IRS and DRS. Because of this unique perspective, DEA (Seiford and Thrall, 1990) proves particularly adept at uncovering such relationship which remain hidden for other methodologies. However, both the methods are in broad agreement as regards the RTS possibilities of the Indian steel industry in the study of Sahoo et al. (1999).

Table 2: Data set for 24 DMUs

DMUs	X1	X2	X3	X4	Y
1	1	1	1	1	1
2	1	1	1	1	2
3	1	1	1	1	3
4	**1**	**1**	**1**	**1**	**4**
5	1	1	2	1	5
6	**1**	**1**	**2**	**1**	**6**
7	1	2	2	1	7
8	**1**	**2**	**2**	**1**	**8**
9	1	2	3	1	9
10	1	2	3	1	10
11	1	2	3	1	11
12	**1**	**2**	**3**	**1**	**12**
13	2	3	4	2	13
14	2	3	4	2	14
15	2	3	4	2	15
16	**2**	**3**	**4**	**2**	**16**
17	2	3	5	2	17
18	**2**	**3**	**5**	**2**	**18**
19	2	4	5	2	19
20	**2**	**4**	**5**	**2**	**20**
21	2	4	6	2	21
22	2	4	6	2	22
23	2	4	6	2	23
24	**2**	**4**	**6**	**2**	**24**

Now let us turn to our data set where some more production possibilities are introduced along with five efficient units[12] (DMUs 4, 6, 8, 12 and 24 in Table 2) of our production process shown in Table 1, and briefly discuss the status of newly introduced DMUs as compared to existing DMUs. The first three DMUs are inefficient because they are producing less output as compared to DMU_4 with a given input combination (1, 1, 1, 1). DMUs 5 and 7 are inefficient as compared to DMUs 6 and 8 respectively. And similar is the case for the case of inefficient DMUs 9, 10, 11 as compared to DMU_{12} with which all the idle capacities existing in various stages are exhausted. However, the remaining DMUs (from DMU 13 until DMU 24) are introduced in the way we have developed the data set from DMU_1 to DMU_{12}. It is to be noted here that if the idle capacities existing in each of

[12] Each production possibility is treated here as a distinct DMU throughout in our analysis, even though they are all one unit only.

the four stages for these DMUs could be taken as the basis for comparison, then the efficiency status of DMUs 16, 18, 20 and 24 could be same as with those of DMUs 4, 6, 8 and 12. *A priori*, we expect these DMUs (all identified with bold letter In Table 2) to operate efficiently because they exhibit two characteristics: not only their outputs are *maximum* with respect to their corresponding given inputs, but also their inputs levels are *minimum*, corresponding with their given outputs. However, the remaining DMUs are inefficient as they produce less as compared to their nearest peers (e.g., DMUS 13, 14, 15 against DMU 16, DMU 17 against DMU 18, DMU 19 against DMU 20, and DMUs 21, 22, 23 against DMU 24).

Table 3: Efficiency scores, slacks, peers and their weights

DMUs	θ	ϕ	f	ι	RTS	Output and Input Slacks*					Peers and their Weights*		
						Q	X1	X2	X3	X4			
1	0.250	1.000	0.250	0.146	IRS	3	0	0	0	0	4 (1)		
2	0.500	1.000	0.500	0.292	IRS	2	0	0	0	0	4 (1)		
3	0.750	1.000	0.750	0.438	IRS	1	0	0	0	0	4 (1)		
4	**1.000**	**1.000**	**1.000**	**0.583**	**CRS**	**0**	**0**	**0**	**0**	**0**	**4 (1)**		
5	0.833	1.000	0.833	0.573	IRS	0	0	0	0.5	0	4 (0.5)	6 (0.5)	
6	**1.000**	**1.000**	**1.000**	**0.688**	**CRS**	**0**	**0**	**0**	**0**	**0**	**6 (1)**		
7	0.875	1.000	0.875	0.656	IRS	0	0	0.625	0.25	0	4 (0.625)	12 (0.375)	
8	1.000	1.000	1.000	0.750	CRS	0	0	0.5	0	0	4 (0.5)	12 (0.5)	
9	0.750	1.000	0.750	0.750	IRS	0	0	0.429	0.643	0	4 (0.214)	12 (0.571)	6 (0.214)
10	0.833	1.000	0.833	0.833	IRS	0	0	0.286	0.429	0	4 (0.143)	12 (0.714)	6 (0.143)
11	0.917	1.000	0.917	0.917	IRS	0	0	0.143	0.214	0	4 (0.071)	12 (0.857)	6 (0.071)
12	**1.000**	**1.000**	**1.000**	**1.000**	**CRS**	**0**	**0**	**0**	**0**	**0**	**12 (1)**		
13	0.812	0.812	0.812	0.655	CRS	0	0.542	0.271	0	0.542	24 (0.083)	12 (0.917)	
14	0.875	0.875	0.875	0.705	CRS	0	0.583	0.292	0	0.583	24 (0.167)	12 (0.833)	
15	0.937	0.937	0.937	0.755	CRS	0	0.625	0.312	0	0.625	24 (0.25)	12 (0.75)	
16	1.000	1.000	1.000	0.806	CRS	0	0.667	0.333	0	0.667	24 (0.333)	12 (0.667)	
17	0.944	0.944	0.944	0.803	CRS	0	0.472	0	0.472	0.472	24 (0.417)	12 (0.583)	
18	1.000	1.000	1.000	0.850	CRS	0	0.5	0	0.5	0.5	24 (0.5)	12 (0.5)	
19	0.950	0.950	0.950	0.831	CRS	0	0.317	0.633	0	0.317	24 (0.583)	12 (0.417)	
20	1.000	1.000	1.000	0.875	CRS	0	0.333	0.667	0	0.333	24 (0.667)	12 (0.333)	
21	0.875	0.875	0.875	0.875	CRS	0	0	0	0	0	24 (0.75)	12 (0.25)	
22	0.917	0.917	0.917	0.917	CRS	0	0	0	0	0	24 (0.833)	12 (0.167)	
23	0.958	0.958	0.958	0.958	CRS	0	0	0	0	0	24 (0.917)	12 (0.083)	
24	**1.000**	**1.000**	**1.000**	**1.000**	**CRS**	**0**	**0**	**0**	**0**	**0**	**24 (1)**		

Note: θ: CCR efficiency score based on the assumption of CRS.
ϕ: BCC efficiency score based on the assumption of VRS.
f: Fare et al. efficiency score based on the assumption of NIRS.
ι: SBM efficiency score based on the assumption of CRS.
* These figures are based on BCC input-oriented model.

We have employed input-oriented CCR, BCC, NIRS and SBM models to compute the efficiency scores of these 24 DMUs, and have used BCC model to account for slacks, peers and their weights. All these results are reported in Table 3. A closer look at the VRS efficiency estimates reveals that only DMUs 4, 6, 12 and 24 are found to be efficient because first, their efficiency scores are each one and second, they do not have any slacks in their corresponding input and output vectors. Prior to the formal programming analysis with DEA, our *ex ante* predictions were that DMUs 4, 6, 8, 12,

16, 18, 20 and 24 were likely to prove efficient under DEA. The results from Table 3 suggest that only four of them, i.e., DMUs 4, 6, 12 and 24 are indeed efficient. This apparent contradiction can probably be explained by slacks alone, which results from the convexity assumption employed in the BCC model, in the estimation of DEA efficiency, because they have all achieved unity efficiency scores. It is interesting to note that all the DMUs expected *a priori* to operate efficiently are deemed efficient in CRS and VRS models (if we take their efficiency scores only). This can be attributed to the failure of both CCR and BCC models declaring them efficient by not accounting for slacks. However, this problem is no longer there in Tone's (2001) SBM measure where only two DMUs[13] (DMU_{12} and DMU_{24}) appear to exhibit full efficiency. This finding has a strong economic implication. These are the only two DMUs who run with their maximum average productivities[14] (or equivalently run with minimum unit costs) where idle capacities are completely exhausted in all the four stages of production.

The CCR, BCC and Färe et al. methods turn some of our *a priori* declared efficient DMUs (8, 16, 18, and 20) into inefficient on the ground that they have slacks in some of their input vectors. To put it differently, had these slacks removed from their inputs, they would have exhibited full efficiency and hence would have operated on the efficient frontier. We find here that these slacks have no economic meaning (why?). Removal of these slacks makes them unable to produce their respective levels of output. This is because the structure of technology is here non-convex (i.e., FDH type), which is again due to the presence of *indivisible* inputs. We then applied the input oriented FDH model to recompute the efficiency scores of these DMUs, and the results are reported in Table 4.

As is seen here, even though all the DMUs have each received unity efficiency score, only our *a priori* declared efficient DMUs are really efficient in terms of 'undomination' criterion. These are the only DMUs that are not dominated by any other DMUs. However, these units dominate the remaining inefficient units, as is seen in the last column of Table 4, declaring themselves as the most dominating units (undominated units) in their corresponding neighborhoods.

[13] In fact, there is only one DMU (DMU_{12}), and the other one, i.e., DMU_{24} is its replica.
[14] In DEA literature, this is called most productive scale size (MPSS).

Table 4: Efficiency scores, slacks, peers and undominated peer: FDH model

DMUs	Eff.[a]	RTS[b]	Output and Input Slacks[a]					Dominating Peers[a]	Most Dominating Peer[a]
			Y	X1	X2	X3	X4		
1	1.000	IRS	3	0	0	0	0	2, 3, 4	4
2	1.000	IRS	2	0	0	0	0	3, 4	4
3	1.000	IRS	1	0	0	0	0	4	4
4	**1.000**	**IRS**	**0**	**0**	**0**	**0**	**0**	**4**	**4**
5	1.000	IRS	1	0	0	0	0	6	6
6	**1.000**	**IRS**	**0**	**0**	**0**	**0**	**0**	**6**	**6**
7	1.000	IRS	1	0	0	0	0	8	8
8	**1.000**	**IRS**	**0**	**0**	**0**	**0**	**0**	**8**	**8**
9	1.000	IRS	3	0	0	0	0	10, 11, 12	12
10	1.000	IRS	2	0	0	0	0	11, 12	12
11	1.000	IRS	1	0	0	0	0	12	12
12	**1.000**	**CRS**	**0**	**0**	**0**	**0**	**0**	**12**	**12**
13	1.000	CRS	3	0	0	0	0	14, 15, 16	16
14	1.000	CRS	2	0	0	0	0	15, 16	16
15	1.000	CRS	1	0	0	0	0	16	16
16	**1.000**	**CRS**	**0**	**0**	**0**	**0**	**0**	**16**	**16**
17	1.000	CRS	1	0	0	0	0	18	18
18	**1.000**	**CRS**	**0**	**0**	**0**	**0**	**0**	**18**	**18**
19	1.000	CRS	1	0	0	0	0	20	20
20	**1.000**	**CRS**	**0**	**0**	**0**	**0**	**0**	**20**	**20**
21	1.000	CRS	3	0	0	0	0	22, 23, 24	24
22	1.000	CRS	2	0	0	0	0	23, 24	24
23	1.000	CRS	1	0	0	0	0	24	24
24	**1.000**	**CRS**	**0**	**0**	**0**	**0**	**0**	**24**	**24**

Note: [a]: These figures are calculated using FDH model.
[b]: RTS figures are calculated using FDH-CRS, FDH-NDRS and FDH-NIRS models.

However, the relevant question now is: can a homogeneous production function estimated by econometric technique exhibit this hidden relationship, i.e., finding out the efficient production possibilities from the technology set and their RTS nature? The answer is generally a negative one because in the absence of an intimate knowledge of the underlying production process, the technology by econometric technique may be *good fit*, but it is otherwise quite arbitrary, whereas the revealed DEA technology is a closer estimate of the true, *unknown* technology underlying the data. So in our above multistage production model, DEA enables us to reveal this hidden relationship by estimating the FDH boundary constituted by the following efficient production possibilities: DMU_4, DMU_6, DMU_8, DMU_{12}, DMU_{16}, DMU_{18}, DMU_{20} and DMU_{24}.

We now examine the question of exploring scale economies in production owing to *process indivisibilities* in this revealed DEA frontier. The scale efficiency index method of Färe et al. (1985) is first employed here to find RTS for each of these DMUs, and the results are reported in RTS column of Table 3. The DMUs, 1, 2, 3, 5, 7, 9, 10, and 11, found to be inefficient in BCC model up to the first 12 units of output of the boundary, are all operating under local IRS, and the

efficient units in this segment are all under local CRS. However, the units operating after 12 units of output of the frontier are all operating under local CRS. We suspect the RTS results of first 12 units (which are contrary to our expectations) revealed from the estimated BCC convex frontier, because the nature of our data suggests the frontier to be non-convex due to indivisibility in the production process. We therefore employed FDH-CRS, FDH-NDRS and FDH-NIRS models to reveal RTS nature along the non-convex frontier. Table 4 (RTS column) reports these results. This frontier reveals the clear trend on RTS possibilities. As expected, DMUs producing below 12 units of output are all characterized by local CRS and DMUs producing 12 or more are under local CRS. As regards the global returns to scale, this revealed technology could be utilized to exploit all the productivity gains due to IRS until all the idle capacities existing at various stages of production are fully exhausted at 12 units of output[15] after which no economies of scale are found. This is also clearly evident in the Figure 3, which is drawn below with given arbitrary factor prices (p_1=10, p_2=7, p_3=5, p_4=6, say).

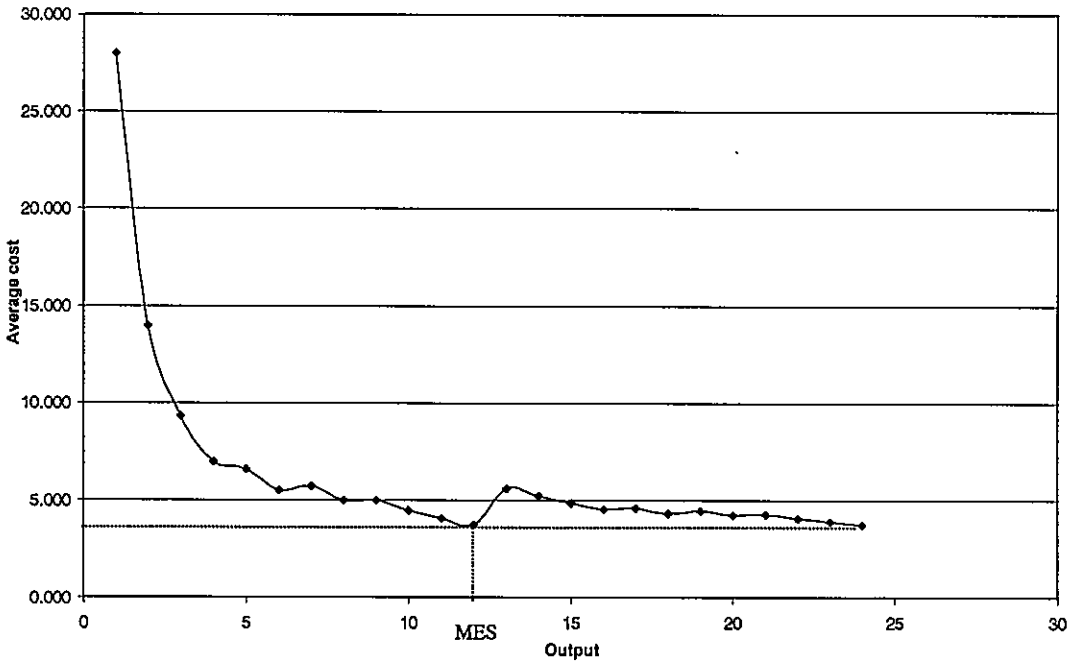


Figure 3: Shape of average cost curve

It is important to observe here that on comparison of the unit cost curves between the first and second sets of 12 DMUs, we see that the extent of steepness of the curve is diminishing. So, as we

[15] In economics literature this scale of output is called minimum efficient scale (MES), where the long-run average cost attains minimum value.

continue generating more and more input-output data points in the way we did for the first and second set of DMUs, we will see that the steepness of the line will tend to vanish and will completely become a completely flat line (i.e., parallel to output axis) when output produced tends to be very large. The implication here is that DMUs operating after MES are all characterized by CRS. These are mere replicas of MES (a scale with zero idle capacities in all its stages).

It is observed from the theory of multi-stage production that *process indivisibilities* arise due to idle capacities. These idle capacities are again due to the unequal task-length of the production runs in the intermediate stages, which make the production technology non-convex as revealed in our constructed DEA boundary, leading to scale economies[16]. However, when parametric technique such as regression is employed to this data set, the resulting estimate will be based on a single optimization over the whole data set, and the fitted technology that results will be an average estimate, which may not replicate the underlying scale behavior of individual DMUs. In employing a series of optimization, one for each firm, DEA provides a *better fit* to each observation and a better approximation to the scale properties of individual firms. As a consequence, the *revealed technology* is a closer estimate of the true, *unknown* technology underlying the data.

It is to be noted here that we have made a clear attempt to explicitly distinguish between these two concepts: returns to scale and economies of scale, which are often used interchangeably in the literature (e.g., Panzar and Willig, 1977, Färe et al., 1988). The former is a characteristic of technology set whereas the latter is associated with a long-run cost function. Even in the case of long-run cost, what are its components? If these components include all costs as highlighted by Classicists, we then feel justified in describing economies of scale as the declining part of long-run AC curve. However, in the special case of given/exogenous input factor prices (what we have maintained while drawing Figure 3), the cost function is entirely determined from an underlying production function where IRS implies economies of scale. However, as the input market is typically imperfect in real world, these two concepts can no longer be same and this distinction warrants caution.

[16] Our result that indivisibility leads to increasing returns gets further support from the firm's equilibrium perspective. A firm facing an imperfectly competitive market attains equilibrium by not operating on CRS facet of the technology. This is because it is the inherent indivisibility (or not all inputs are taken into account) in the production process that forces the firm to operate under IRS. See Appendix B for a detailed discussion on it.

The next section describes the actual production process in cement manufacturing. The presence of scale economies due to *process indivisibilities* is clearly demonstrated by this description.

4. Towards an empirical application

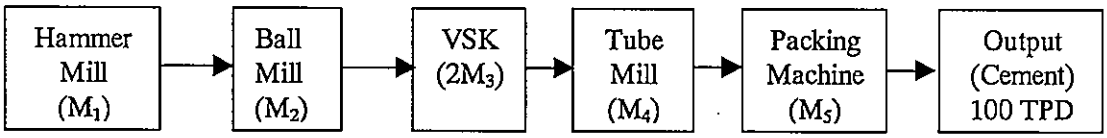
The cement manufacturing firm is taken here as an example to show how economies of scale in production arise mainly due to *technique* as well as *indivisibilities in process*. We have shown here two representative mini-cement plants (out of five plants) of varying capacities. The techniques used in this industry are of two types: Vertical Shaft Kiln (VSK) [capacity: 50 tones per day (TPD)] and Rotary Kiln [capacity: 200 TPD]. The data are collected from the funding agency, Andhra Pradesh Industrial Development Corporation (APIDC), Hyderabad, India. The difference between the two techniques is one of the important sources of scale in cement manufacturing. The main piece of capital equipment that differentiates the two techniques is the kiln in which a rotary feeder distributes uniformly over the entire cross-section of the fire bed. Table 5 shows the main plant and machinery of one of the representative plants.

Table 5: Main plant and machinery: vertical shaft kiln (Technique 1)

Sl. No.	Department (Stages)	Specification (Equipment)	Capacity
1	Lime Stone Crushing	Hammer Mill	50 TPH
2	Raw Mill	Ball Mill	11 TPH
3	Kiln Section	Vertical Shaft Kiln (2 Nos.)	50 TPD each.
4	Cement Mill	Tube Mill	7 TPH
5	Packing House	Single Spout (Packing Machine)	10 TPH

Note: TPH: Tones per hour.

In terms of our presentation of production process, it looks like:



Here the total production process[17] is divided into five principle stages, and the task-lengths associated with each of these stages are not equal. At the end of the workday, 100 tones of cement are produced with idle capacities existing in all the stages excepting at Stage 3 (Kiln Section). However, in order to meet the increase in demand, this plant has actually increased its production

[17] For a brief description of the production process in cement manufacturing, see Appendix C.

to 250 TPD by adding three vertical shaft kilns to the existing line of production, which has resulted a fall in the unit cost of production. But, further production (above 250 TPD) by adding more VSK to the existing line is not technically feasible because this additional increase in output requires not only the addition of VSKs but also some civil works, i.e., Kiln house structure, kiln bed foundation, raw mill foundation, and clinker storage yard have to be reconstructed to accomplish this further production, all of which requires some additional cost. However, a consultation with Deputy General Manager of APIDC reveals that it will be cost effective if the other technique, Rotary Kiln[18] is adopted at the capacity level of 200 TPD. Even though the cost of Rotary Kiln is higher than that of VSK, the cost of civil works is much more than this price difference between Rotary kiln and VSK. Also, some plants that have used Rotary Kiln (Technique 2) have expanded their production up to 600 TPD just by mere adding two more rotary kilns to their existing line and have also experienced a decline in unit cost. So what we observe here is that unit cost of production falls due to two reasons: 1) differential increase in some/all inputs, which are again due to unequal task-lengths associated with various stages of production, and 2) better technique, which is cost efficient at the higher stage of production.

Table 6: Main plant and machinery: rotary kiln (Technique 2)

Sl. No.	Department (Stages)	Specification (Equipment)	Capacity
1	Lime Stone Crushing	Impact Crusher (1000x1000m)	100 TPH
2	Raw Mill	2.8m. dia. x 7.5m. long	25 TPH
3	Rotary Kiln	3m. dia. x 45m. long	200TPD each.
4	Cooler	1.8m x 8.8m.	200 TPD
5	Cement Mill	2.4m. dia. x 10m. long	15 TPH
6	Coal Mill	2.2m. dia. x 6m. long	5 TPH
7	Packing	Single Spout (Packing Machine)	40 TPH

5. Summary and conclusion

We have discussed in depth the historical evolution of scale since Adam Smith and have brought the distinction between two concepts: returns to scale and economies of scale by relating the former to the concept of 'production unit' and the latter to the concept of 'firm' and finally arrive at the conclusion that the former is a component of the latter. The reason for these two terms being used interchangeably in the literature is that the neoclassicists are preoccupied with homogeneous functional form for the production function to describe returns to scale. Here the requirement of

[18] This technique, however, is not efficient at a smaller stage of production, e.g., at 50 or 100 TPD.

equiproportionate changes in factors was not aimed at refining the concept of scale but to develop a 'theory of distribution' under competitive market structure.

We then develop a simple multi-stage model of a production process, and show how *process indivisibilities* arise and how they could lead to scale effects. In a multistage production process idle capacity may arise due to unequal length of production runs of intermediate stages, which leads to scale effects when production is expanded. If final output can be scaled to be nearest integer value of that production run which has the largest idle capacity, then economies of scale are realized since total costs do not increase proportionately to the volume of output. Such a characteristic is called *process indivisibility* and would be a common feature in almost all the multi-stage production processes. The relevant question that is posed now is: can a homogeneous characterization of production function capture scale if it arises in this fashion? The answer to the question is generally a negative one. However, it is argued that this inability of the production function to capture scale arising from such sources is not because the notion of production function precludes the incorporation of such features; rather it is the homogeneous property of the production function that leads us astray. We have shown here that the non-convex FDH technology in the multi-stage production model reveals non-homogeneity and discreteness in character; and captures scale effects arising from *process indivisibilities*. However, the standard convex nonparametric technologies embedded in BCC and CCR models fail to clearly exhibit such scale effects.

Implications to managers and academicians

Since most of the business entities are faced with intense competition, the only way to survive and prosper for a unit is to constantly improve its relative performance in the industry. One way is to expand production to operate at full capacity unless the market can be served with one unit of the output operating at less than full capacity. In other words, economies of scale owing to all sources (including process indivisibility) need to be fully exploited till MES is reached. DEA enables the manager to obtain such unit specific information on RTS possibilities as well as MES. Further, this piece of information also helps in indicating potential redistribution of resources among firms through mergers and acquisitions.

To the defense that the neoclassical production function is a toolkit that can be used to study the RTS behavior of the business entities in the industry, one needs the further reinterpretation of

Koopmans' proportionality postulate. As the proportionality postulate itself stands, it obscures countless scale effects because of its high level of abstraction. As we have argued earlier, the interpretation of $\lambda.K$ is not that λ times K but the volume of capital embodied in $\lambda.K$. And similar reinterpretation for labor also holds true. Otherwise, the neoclassical production function will always exhibit CRS, assuming away all possible relevant scale effects actually operating in the plant. Most of the existing DEA models that are used to provide information on RTS possibilities obscure economic dimensions. We have made an attempt here by exploring the indivisibility dimension in FDH model as a possible source of scale economies. We do expect future DEA researchers to explore other economic dimensions of returns to scale (as has been expounded by Classicists) in the current existing DEA models, which will serve to bridge up the significant divergences between econometric and DEA approaches for the estimation of production frontier.

References

1. Allport, J. A. and C. M. N. Stewert, 1978, Economics (Cambridge University Press).

2. Aoki, M., 1990, The firm as a nexus of treaties (ed.), (Sage Publication, London).

3. Arrow, K. J., H. B. Chenery, B. S. Minhas and R. M. Solow, 1961, Capital-labor substitution and economic efficiency, Review of Economics and Statistics 53, 225-251.

4. Banker, R. D., 1984, Estimation of most productive scale size, European Journal of Operational Research 17, 35-44.

5. Banker, R. D., I. Bardhan and W. W. Cooper, 1996a, A note on returns to scale in DEA, European Journal of Operational Research 88, 583-585.

6. Banker, R. D., H. Chang, and W. W. Cooper, 1996b, Equivalence and implementation of alternative methods for determining returns to scale in data envelopment analysis, European Journal of Operational Research 89, 473-481.

7. Banker, R. D., A. Charnes and W. W. Cooper, 1984, Some models for estimating technical and scale inefficiencies in data envelopment analysis, Management Science 30, 1078-1092.

8. Banker, R. D., A. Charnes, W. W. Cooper, J. Swarts and D. Thomas, 1989, An introduction to data envelopment analysis with some of its models and their uses, Research in Government and Non-Profit Accounting 5, 125-163.

9. Banker, R. D., R. F. Conrad and R. P. Strauss, 1986, A comparative application of data envelopment analysis and translog methods: an illustrative study of hospital production, Management Science 32, 30-44.

10. Banker, R. D. and R. M. Thrall, 1992, Estimation of returns to scale using data envelopment analysis, European Journal of Operational Research 62, 74-84.

11. Chamberlin, E., 1947-48, Proportionality, divisibility and economies of scale, Quarterly Journal of Economics 62(b), 229-262.

12. Chamberlin, E., 1948, The theory of monopolistic competition (Harvard University Press, Cambridge).

13. Charnes, A., W. W. Cooper and E. Rhodes, 1978, Measuring the efficiency of decision making units, European Journal of Operational Research 2, 429-444.

14. Chenery, H. B., 1949, Engineering production functions, Quarterly Journal of Economics 63, 501-531.

15. Clark, J. M., 1923, Studies in the economies of overhead costs (Chicago University Press, Chicago).

16. Cooper, W. W., L. M. Seiford and K. Tone, 1999, *Data envelopment analysis: a comprehensive text with models, applications, references and DEA-solver software* (Kluwer Academic Publishers, Boston).

17. Douglas, P. H., 1948, Are there laws of production?, American Economic Review 38, 1-41.

18. Färe, R. and S. Grosskopf, 1994, Estimation of returns to scale using data envelopment analysis: a comment, European Journal of Operational Research 79, 379-382.

19. Färe, R., S. Grosskopf and C. A. K. Lovell, 1985, The measurement of efficiency of production (Kluwer Nijhoff, Boston).

20. Färe, R., S. Grosskopf and C. A. K. Lovell, 1988, Scale elasticity and scale efficiency, Journal of Institutional and Theoretical Economics 144, 721-729.

21. Golany, B. and G. Yu, 1997, Estimating returns to scale in DEA, European Journal of Operational Research 103, 28-37.

22. Gold, B., 1981, Changing perspectives on size, scale and returns: an interpretative survey, Journal of Economic Literature 19, 5-33.

23. Haldi, J. and D. Whitcomb, 1967, Economies of scale in industrial plants, Journal of Political Economy 75, 373-385.

24. Kaldor, N., 1934, The equilibrium of the firm, Economic Journal 34, 60-76.

25. Kerstens, K. and P. Vanden Eeckaut, 1999, Estimating returns to scale using non-parametric technologies: a new method based on goodness-of-fit, European Journal of Operational Research 113, 206-214.

26. Koopmans, T. C., 1957, Three essays on the state of economic science (McGraw Hill, New York).

27. Marshall, A., 1920, Principles of economics (Macmillan, London).

28. Marx, K., 1978, Capital (Penguin Books).

29. Panzar, W. and R. D. Willig, 1977, Economies of scale in multi-output production, Quarterly Journal of Economics 91, 481-493.

30. Robinson, E. A. G., 1935, The structure of competitive industry (Pitman, New York).

31. Robinson, J., 1969, The economics of imperfect competition (Macmillan, London).

32. Russell, R. R. and M. Wilkinson, 1979, Microeconomics: a synthesis of modern and neoclassical theory (John Wiley, New York).

33. Sahoo, B. K., P. K. J. Mohapatra and M. L. Trivedi, 1999, A comparative application of data envelopment analysis and frontier translog production function for estimating returns to scale and efficiencies, International Journal of Systems Science 30, 379-394.

34. Samuelson, P. A., 1947/65, Foundations of economic analysis (Atheneum, New York).

35. Samuelson, P. A., 1961, Economics: an introductory analysis (McGraw-Hill, New York).

36. Seiford, L. M. and R. Thrall, 1990, Recent developments in DEA: The mathematical programming approach to frontier analysis, Journal of Econometrics 46, 7-38.

37. Sengupta, J. K., 1987, Efficiency measurement in non-market systems through data envelopment analysis, International Journal of Systems Science 18, 2279-2304.

38. Silberston, Z. A., 1972, Economies of scale in theory and practice, Economic Journal 82, 369-391.

39. Smith, A., 1791, An inquiry into the wealth of nations (Strahan and Cadell, London).

40. Soni, K. C. and B. B. Jani, 1987, VES production function with neutral technological change: a comparative study of the industrial sectors of Gujarat state vs. all India, Anvesak 17, 59-75.

41. Stigler, G. J., 1946, Production and distribution theories: The formative period (Macmillan, New York).

42. Stiglitz, J., 1975, Information and economic analysis, in: M. Parkan and A. Nobay, eds., Current economic problems (Cambridge University Press, Cambridge).

43. Tone, K., 1996, A simple characterization of returns to scale in DEA, Journal of the Operations Research Society of Japan 39, 604-613.

44. Tone, K., 2001, On returns to scale under weight restrictions in data envelopment analysis, Journal of Productivity Analysis 16, 31-47.

45. Tone, K., 2001, A slacks-based measure of efficiency in data envelopment analysis, European Journal of Operational Research 130, 498-509.

46. Tulken, H. and P. Vanden Eeckaut, 1995, Non-parametric efficiency, progress and regress measures for panel data: methodological aspects, European Journal of Operational Research 80, 474-499.

47. Varian, H. R., 1992, Microeconomic analysis (W. W. Norton & Company, New York).

Appendix A: Derivation of cost function from C-D production function

$Min_{L,K} \quad C = P_L.L + P_K.K$

$s.t. \quad Y = A.L^{\alpha}.K^{\beta}$

Now let us formulate the lagranzian function as follows:

$Z = P_L.L + P_K.K + \lambda.[Y - A.L^{\alpha}.K^{\beta}]$

$$\frac{\partial Z}{\partial L} = P_L - \lambda.\alpha.A.L^{\alpha-1}.K^{\beta} = 0 \qquad \text{.....(A1)}$$

$$\frac{\partial Z}{\partial K} = P_K - \lambda.\beta.A.L^{\alpha}.K^{\beta-1} = 0 \qquad \text{.....(A2)}$$

$$\frac{\partial Z}{\partial \lambda} = Y - A.L^{\alpha}.K^{\beta} = 0 \qquad \text{.....(A3)}$$

From Eqn. (A1), $\lambda = \frac{P_L.L}{\alpha.Y}$, and from Eqn. (A2), $\lambda = \frac{P_K.K}{\beta.Y}$. So,

$$\lambda = \frac{P_L.L}{\alpha.Y} = \frac{P_K.K}{\beta.Y}$$

$$or, \quad K = \left(\frac{\beta}{\alpha}\right).\left(\frac{P_L}{P_K}\right).L$$

Now, on substitution of the optimal value of K in total cost function leads to

$$C = P_L.L + P_K.\left(\frac{\beta}{\alpha}\right).\left(\frac{P_L}{P_K}\right)L = P_L.L\left(1+\frac{\beta}{\alpha}\right) \qquad \text{.....(A4)}$$

From Eqn. (A3), we get $Y = A.L^{\alpha}.K^{\beta}$. Substituting the value of the optimal K in this production function leads us to express L in terms of Y as follows:

$$L = \left(\frac{Y}{A}\right)^{\frac{1}{\alpha+\beta}}.\left[\left(\frac{\alpha}{\beta}\right).\left(\frac{P_K}{P_L}\right)\right]^{\frac{\beta}{\alpha+\beta}}$$

Substituting this value of L in Eqn. (A4) yields the following cost function

$$C = \left(\frac{\alpha+\beta}{\alpha}\right).\left[\frac{1}{A}\left(\frac{\alpha}{\beta}\right)^{\beta}.P_L^{\alpha}.P_K^{\beta}\right]^{\frac{1}{\alpha+\beta}}.Y^{\frac{1}{\alpha+\beta}}$$

$$= \bar{C}.Y^{\frac{1}{\alpha+\beta}}, where \quad \bar{C} = \left(\frac{\alpha+\beta}{\alpha}\right).\left[\frac{1}{A}\left(\frac{\alpha}{\beta}\right)^{\beta}.P_L^{\alpha}.P_K^{\beta}\right]^{\frac{1}{\alpha+\beta}}.$$

Appendix B

The production function must be homogeneous of degree one of all the variables, and if this is not so, it must be either because of 'indivisibility' or because not all inputs have been taken into account (Samuelson, 1947/65, pp.84-85).

Following Robinson (1969), 'equilibrium' is defined in an imperfectly competitive market as an output level where first, marginal revenue (MR) equals marginal cost (MC) (because profit maximization is assumed) and second, where total revenue (TR) equals total cost (TC) (because competition is assumed to be sufficient enough to completely eliminate excess profits). From this definition we see that all producers are necessarily producing at a level of output for which the average cost (and the price) is above the possible minimum. Stiglitz (1975), among others, argues that the imperfectly competitive equilibrium is sub-optimal. In one sense, the firm in this equilibrium is facing increasing returns, since $MC < AC$. If the firm's AC curve accounts for costs of all inputs, then it cannot be maximizing with respect to all inputs. Let us elaborate it further.

Let us consider Samuelson's (1947/65, pp. 84-85) example of a simple economy where labor (L) and capital (K) are truly the only inputs into the production of good Q, where the production function, $Q = f(K, L)$, for a firm in an imperfectly competitive market is formally assumed to be linearly homogeneous. Then, according to Euler's Theorem, the following holds true:

$$Q = MP_K * K + MP_L * L \qquad \text{.....(B1)}$$

where, MP_L and MP_K are marginal product of labor and capital respectively. Now consider the relationship between price of output (P) (which is also called average revenue (AR)), MR and price elasticity of demand (η_p).

$$MR = P * \left(1 + \frac{1}{\eta_p}\right)$$

At equilibrium, $MR = MC$, so

$$MC = P * \left(1 + \frac{1}{\eta_p}\right), or, P = MC / \left(1 + \frac{1}{\eta_p}\right)$$

The short-run MC depends upon which input is being considered variable to calculate MC. Assuming L to be variable; MC is then the cost of the additional units of labor required to produce an additional unit of output:

$$MC=P_L/MP_L$$

When the imperfectly competitive firm is maximizing profit with respect to labor, then

$$MP_L=(P_L/P)/\left(1+\frac{1}{\eta_p}\right) \qquad \text{.....(B2a)}$$

Similarly, in case of capital, if considered being variable,

$$MP_K=(P_K/P)/\left(1+\frac{1}{\eta_p}\right) \qquad \text{.....(B2b)}$$

The second condition for imperfectly competitive equilibrium requires TR to equal to TC; then the following equation holds good.

$$P*Q=P_L*L+P_K*K$$
$$or,Q=(P_L/P)*L+(P_K/P)*K \qquad \text{.....(B3)}$$

Substituting (B2a) and (B2b) into (B3) yields the following:

$$Q=\left(1+\frac{1}{\eta_p}\right)*\left[MP_L*L+MP_K*K\right] \qquad \text{.....(B4)}$$

Let us now compare these two equations: (B1) and (B4). If $(1/\eta_p) \neq 0$ in (B4), then the production function for profit maximizing imperfect competitor in equilibrium must not homogeneous with respect to L and K alone. Furthermore, it must exhibit increasing returns to scale if (B4) is also to be true. But if (B4) is true, then (B1) cannot be true. If (B1) is true, the excess profits are not zero, or are not being maximized with respect to all inputs.

If (B1) does not hold, it could be that not all inputs are truly variable, or that not all inputs are recognized (Samuelson, 1947/65, pp. 84-85). In either case, it means that there is some constraint (what we refer to it here as *process indivisibility*) on the production function, which is distorting the usual equilibrium results.

Appendix C: A brief description of the production process

Limestone from the quarries will be transported by trucks and stored in the storage hopper and conveyed to Hammer Mill and crushed to minus 15 mm. This final crushed limestone carried through an inclined belt conveyer to the storage yard. These additives as per raw mix design and fuel will be stored in respective storage yard.

The ground material will be drawn by table feeders as per the raw mix design to the grinding mill by the inclined belt conveyer to grind the raw mix to the required fineness and then transported to the blending silos with the sucking fan. Pneumatic blending is carried out in both the blending-cum-storage silos. The blending unit will be provided with bag filter for dust emission.

The raw mill from the storage silos is extracted by rotary feeders and then transported to the raw mill hoper by means of a screw conveyor and vertical pneumatic gravity conveyor (Air Lift). The raw mill feed rate to the noduliser is also controlled. The nodules from the nodulisers are fed into the vertical shaft kiln through a damper control valve. A rotary feeder distributes the nodules uniformly over the entire cross section of the fire bed in the kiln. These nodules then travel downwards in the kiln and undergo various reactions and get converted into clinker. The air will be supplied through a root blower. The hot exhaust gases escape through the chimney. The clinker is discharged through rotary grate and transported to the clinker storage shed by means of belt conveyor after passing through the jaw crusher to reduce the size of the clinker lumps. Clinker is stored in a storage yard. Clinker and gypsum are transported by means of belt conveyor and fed through the table feeders to the cement-grinding mill. The cement mill is an open circuit, three-component tube mill. The ground cement is transported to the cement silos by means of vertical pneumatic gravity conveyor. The cement is extracted from the cement storage silo through a rotary feeder and rotary screen and packed in bags by a single spout cement-packing machine. The spilled over cement is collected in a hopper and recycled again through a screw conveyor.