Scale, Indivisibilities and Production Function in Data Envelopment Analysis

Kaoru Tone*and Biresh K. Sahoo [†]
National Graduate Institute for Policy Studies [‡]

1. Introduction

The term 'economies of scale' is defined in the literature in two alternative ways: either in terms of physical output or cost of production. The neoclassical 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 neoclassical 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. 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

*Corresponding author. tone@grips.ac.jp

[†] biresh@grips.ac.jp

[‡] 2-2 Wakamatsu-cho, Shinjuku-ku, Tokyo 162-8677, Japan.

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 pointing 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 (DEA), originally developed by Charnes *et al.* (1978), 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 deals with the implications of this study to managers and academicians.

2. Historical evolution of scale

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 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.

Some authors such as Russell and Wilkinson (1979), 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 nonlinear. 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 , 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.

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.

Indivisibility argument to explanaing 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 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.

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.

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.

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.

3. Multistage production model

We 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 multi-stage 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 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 the 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 (Kerstens and Vanden Eeckaut, 1999) 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.

4. Towards an empirical application

The cement manufacturing firm is taken here as an example to show how scale economies in production arise mainly due to technique as well as process *indivisibilities.* We have shown here two representative mini-cement plants (out of five) 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.

Here the total production process 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 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 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 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 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.

5. Implications

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 λK is not that λ times K but the volume of capital embodied in λK. 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] Aoki, M. (1990) *The firm as a nexus of treaties* (ed.), (Sage Publication, London).

[2] Chamberlin, E. (1947-48) "Proportionality, divisibility and economies of scale," *Quarterly Journal of Economics*, 62, 229-262.

[3] Charnes, A., W. W. Cooper and E. Rhodes (1978) "Measuring the efficiency of decision making units," *European Journal of Operational Research*, 2, 429-444.

[4] Clark, J. M. (1923) *Studies in the economies of overhead costs*, (Chicago University Press, Chicago).

[5] Kaldor, N. (1934) "The equilibrium of the firm," *Economic Journal*, 34, 60-76.

[6] 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.

[7] Marshall, A. (1920) *Principles of economics*, (Macmillan, London).

[8] Marx, K. (1978) *Capital*, (Penguin Books).

[9] Robinson, E. A. G. (1935) *The structure of competitive industry*, (Pitman, New York).

[10] Robinson, J. (1969) *The economics of imperfect competition*, (Macmillan, London).

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

[12] Silberston, Z. A. (1972) "Economies of scale in theory and practice," *Economic Journal*, 82, 369-391.

[13] Smith, A.(1791) *An inquiry into the wealth of nations*, (Strahan and Cadell, London).