Degree of Scale Economies and Congestion: A Unified DEA Approach

01302170 National Graduate Institute for Policy Studies
Kaoru TONE* and Biresh K SAHOO

1. Introduction

Most of the business units (or decision making units (DMUs)) today face an irritatingly limited supply of resources and face persistent competition. This has led to a significant emphasis on the efficient utilization and allocation of on-hand resources by building larger operating units to achieve the possible advantages of 'scale economies.' From a policy point of view, the estimation of scale elasticity (returns to scale) parameter is of particular importance concerning whether there is any scope for increased productivity by expanding/contracting, whether minimum efficient scales will allow competitive markets to be established, and if the existing size distribution of firms is consistent with a competitive market outcome.

There are many resources which affect the performance of a unit when there is 'overuse' of such resources. When firms use these resources they only take into account their own benefits and costs of such usage, but largely ignore the congestion, or exclusion costs that they impose on others. This is referred as "congestion externality" in economics literature. When congestion is present, it effectively shrinks business market areas and reduces the scale economies. So there is a need to estimate the returns to scale parameter in the presence of congestion, and to examine it with more prudence for the firm's financial viability and success.

Since the breakthrough by Banker, Charnes and Cooper (1984), the returns to scale issues under multiple input/output environments have been extensively studied in the framework of DEA. The treatment of congestion within the DEA framework has received considerable attention in the recent literature. After the concept of congestion was first introduced in the paper by Färe and Svensson (1980), subsequently was given operationally implementable form by Färe *et al.* (1985) and Cooper *et al.* (1996, 2000), there has been found a growing interest in congestion in a number of application areas, e.g., congestion in Chinese production by Brockett *et al.* (1998), congestion in US teaching hospitals by Grosskopf *et al.* (1998). In the presence of congestion, however, the studies on estimation of scale elasticity are almost nil. This paper makes a humble attempt to fill in this void.

2. Scale Elasticity in Production

Throughout this paper, we deal with n DMUs, each having m inputs for producing s outputs. For each DMU$_o$ $(o = 1, \ldots, n)$, we denote respectively the input/output vectors by $\boldsymbol{x}_o \in R^m$ and $\boldsymbol{y}_o \in R^s$. The input/output matrices are defined by $X = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n) \in R^{m\times n}$ and $Y = (\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n) \in R^{s\times n}$. We assume that $X > O$ and $Y > O$.

The returns to scale (RTS) or scale elasticity in production (ρ) or degree of scale economies (DSE) is defined as the ratio of marginal product (MP) to average product (AP). In a single input/output case, if the output y is produced by the input x, we define the production elasticity ρ by

$$\rho = \mathrm{MP}/\mathrm{AP} = \frac{dy}{dx}\Big/\frac{y}{x}. \quad (1)$$

RTS is said to be *increasing, constant* and *decreasing* if $\rho > 1$, $\rho = 1$ and $\rho < 1$ respectively (See Baumol *et al.* (1988)).

We deal with the production possibility set P_{BCC} related with the multiple input-output correspondence as defined by

$$P_{BCC} = \{(\boldsymbol{x}, \boldsymbol{y}) | \boldsymbol{x} \geq X\boldsymbol{\lambda}, \boldsymbol{y} \leq Y\boldsymbol{\lambda}, \boldsymbol{e}\boldsymbol{\lambda} = 1, \boldsymbol{\lambda} \geq \boldsymbol{0}\}, \quad (2)$$

where $\boldsymbol{e}$ is a row vector with all its elements being equal to one.

Analogous to the single input-output case, we measure the production elasticity of DMUs positioned on the efficient portion of P_{BCC}.

The dual program to the [BCC-O] model is described as follows, using the dual variables $\boldsymbol{v} \in R^m$, $\boldsymbol{u} \in R^s$ and $w \in R$:

$$\begin{aligned} &[\text{Dual}] && \min \boldsymbol{v}\boldsymbol{x}_o - w && (3)\\ &\text{subject to} && -\boldsymbol{v}X + \boldsymbol{u}Y + \boldsymbol{e}w \leq \boldsymbol{0} && (4)\\ & && \boldsymbol{u}\boldsymbol{y}_o = 1 && (5)\\ & && \boldsymbol{v} \geq \boldsymbol{0},\ \boldsymbol{u} \geq \boldsymbol{0}. && (6)\end{aligned}$$

If we assume that the DMU $(\boldsymbol{x}_o, \boldsymbol{y}_o)$ is (strongly) efficient, there exists an optimal solution $(\boldsymbol{v}^*, \boldsymbol{u}^*, w^*)$ for [Dual] such that:

$$\boldsymbol{v}^* \boldsymbol{x}_o - w^* = 1. \tag{7}$$

The hyperplane $-\boldsymbol{v}^*\boldsymbol{x} + \boldsymbol{u}^*\boldsymbol{y} + w^* = 0$ is a supporting hyperplane to the production possibility set P_{BCC} at $(\boldsymbol{x}_o, \boldsymbol{y}_o)$.

We now define respectively the virtual input (ξ) and virtual output (η) associated with $(\boldsymbol{x}, \boldsymbol{y}) \in P_{BCC}$ by

$$\xi = \boldsymbol{v}^*\boldsymbol{x} \quad \text{and} \quad \eta = \boldsymbol{u}^*\boldsymbol{y}. \tag{8}$$

Thus, the hyperplane can be expressed as:

$$-\xi + \eta + w^* = 0. \tag{9}$$

From this equation, we derive marginal product (MP) as:

$$\text{MP} = \frac{d\eta}{d\xi} = 1, \tag{10}$$

and average product (AP) as:

$$\text{AP} = \frac{\eta}{\xi} = \frac{\eta}{\eta + w^*}. \tag{11}$$

At $(\boldsymbol{x}_o, \boldsymbol{y}_o)$, we have $\eta = 1$, and hence the production elasticity (ρ) or DSE at $(\boldsymbol{x}_o, \boldsymbol{y}_o)$ is expressed by:

$$\rho = 1 + w^*. \tag{12}$$

We also solve:

$$
\begin{aligned}
&[\text{Upper(Lower)}] \quad && \bar{w}(\underline{w}) = \max(\min) w && (13)\\
&\text{subject to} && -\boldsymbol{v}X + \boldsymbol{u}Y + \boldsymbol{e}w \leq \boldsymbol{0} && (14)\\
& && -\boldsymbol{v}\boldsymbol{x}_o + \boldsymbol{u}\boldsymbol{y}_o + w = 0 && (15)\\
& && \boldsymbol{u}\boldsymbol{y}_o = 1 && (16)\\
& && \boldsymbol{v} \geq \boldsymbol{0},\ \boldsymbol{u} \geq \boldsymbol{0}. && (17)
\end{aligned}
$$

The upper (lower) scale elasticity in production $\bar{\rho}$ $(\underline{\rho})$ is calculated by:

$$\bar{\rho} = 1 + \bar{w} \quad \text{and} \quad \underline{\rho} = 1 + \underline{w}. \tag{18}$$

3. Congestion

So far we have dealt with situations where input slacks (excesses) are considered free. The set P_{BCC} allows an (unbounded) input $\boldsymbol{x} \geq X\boldsymbol{\lambda}$ for producing an output $\boldsymbol{y} = Y\boldsymbol{\lambda}$. Under this assumption, the scale elasticity ρ is nonnegative. However, there are some cases in which an increase in one or more inputs causes the worsening of one or more outputs. A typical example is the case of mining where too many miners in an underground mine may lead to "congestion" with reference to output. In order to potentially deal with such situation, we need to modify our production possibility set as follows:

$$P_{convex} = \{(\boldsymbol{x}, \boldsymbol{y}) | \boldsymbol{x} = X\boldsymbol{\lambda}, \boldsymbol{y} \leq Y\boldsymbol{\lambda}, \boldsymbol{e}\boldsymbol{\lambda} = 1, \boldsymbol{\lambda} \geq \boldsymbol{0}\}. \tag{19}$$

We discuss the scale elasticity issue with respect to this new production possibility set P_{convex}, and demonstrate that "congestion" is recognized by the status with having negative production elasticity $(\rho < 0)$, and that the degree of congestion can be measured by ρ.

4. Weak congestion

The common understanding on congestion is that an increase (decrease) in one or more inputs causes a decrease (increase) in one or more outputs (Cooper *et al.*, 2001).

Definition 1 *(Weak congestion)*
A DMU is (weakly) congested if it is strongly efficient with respect to P_{convex} and there exists an activity in P_{convex} that uses less resources in one or more inputs for making more products in one or more outputs.

We have the following theorem regarding the status of weak congestion and inefficiency in the BCC-O model.

Theorem 1 *Suppose that the DMU $(\boldsymbol{x}_o, \boldsymbol{y}_o)$ is efficient with respect to the P_{convex}. Then, it is weakly congested if and only if it has $\theta^*_{BCC} > 1$ or $(\theta^*_{BCC} = 1$ and $\boldsymbol{s}^{+*} \neq \boldsymbol{0})$ by the model* [BCC-O].

5. An empirical study

We will demonstrate a case study regarding the operations of supermarkets in Japan on site.

6. Conclusion

Investigation of scale elasticity for obtaining optimal scale of operations has significant bearings while recommending policy for restructuring any sector in a competitive economy. In recent years the BCC model has enjoyed widespread popularity in the non-parametric literature for computing the scale elasticity in production. However, there is a difficulty with the use of BCC model since this model does not take the congestion factors into consideration. When congestion is present in production, BCC model overstates true scale elasticity estimates. It is of interest, then, to determine the impact of congestion on scale elasticity parameter of incorporating congested production factors into the model. This paper makes a contribution to the literature in this regard by proposing a new method to measure the scale elasticity parameter in the presence of congestion.