GRIPS Research Report Series I-2002-0005

Degree of Scale Economies and Congestion: A Unified DEA Approach

By

Kaoru Tone and Biresh K. Sahoo

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

First version: August 07, 2002

Research supported by Grant-in-Aid for Scientific Research(C)
Japan Society For the Promotion of Science

GRIPS
NATIONAL GRADUATE INSTITUTE
FOR POLICY STUDIES

Degree of Scale Economies and Congestion: A Unified DEA Approach

By

Kaoru Tone and Biresh K. Sahoo

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

First version: August 07, 2002

Research supported by Grant-in-Aid for Scientific Research (C)
Japan Society for the Promotion of Science

Degree of Scale Economies and Congestion: A Unified DEA Approach

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

Abstract

There are increasing concerns about how increase in congestion can adversely affect output as well as about the relative benefit-cost ratio or return on investment associated with alternative projects or policies to address those problems. Regardless of what policy strategies are used to address congestion, the fact remains that we can not assess the economic benefits of congestion-reduction strategies unless we are able to measure the extent to which congestion affects productivity in general, and scale economies in particular. This paper makes a novel attempt to suggest a method in a non-parametric framework to measure scale elasticity in production in the presence of congestion.

Keywords: DEA, scale elasticity, degree of scale economies, congestion.

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

*tone@grips.ac.jp
†biresh@grips.ac.jp
‡2-2 Wakamatsu-cho, Shinjuku-ku, Tokyo 162-8677, Japan.

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 (Data Envelopment Analysis). 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), and 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.* (2001). Just as the DEA literature (Färe *et al.*, 1986, 1988, Banker *et al.*, 1996a,b, Førsund, 1996 and Sueyoshi, 1997, 1999) that addresses the evaluation of scale elasticity is sparse, the research on using DEA to evaluate scale elasticity in the presence of congestion is almost nil. This paper makes a humble attempt to fill in this void.

This paper unfolds as follows. Section 2 discusses the scale elasticity or the degree of scale economies (DSE). Then we turn our attention to the problem of scale elasticity issue in the presence of congestion, and propose a new method for identifying DSE in Section 3. Sections 4 and 5 introduce an illustrative example and an empirical case study regarding the scale elasticity of operations of supermarkets of the Japan Chain Stores Association. Some concluding remarks follow in Section 6.

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) or *Passus Coefficient*, 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 scale elasticity ρ by

$$\rho = \text{MP}/\text{AP} = \frac{dy}{dx} \Big/ \frac{y}{x}. \tag{1}$$

See Hanoch (1970), Starrett (1977), Panzar and Willig (1977) and Baumol *et al.* (1988) for the detailed discussion.

For a neoclassical 'S-shaped production function' (or *Regular Ultra Passum Law* (RUPL) in the words of Frisch, 1965), ρ can take on values ranging from 'greater than one' to 'less than one,' and even negative values when production decreases with usage of inputs. RTS is said to be *increasing, constant* and *decreasing* if $\rho > 1$, $\rho = 1$ and $\rho < 1$ respectively. The production function $y = f(x)$ satisfies RUPL if $\partial\rho/\partial y < 0$ and $\partial\rho/\partial x < 0$ (Førsund and Hjalmarsson, 2002).

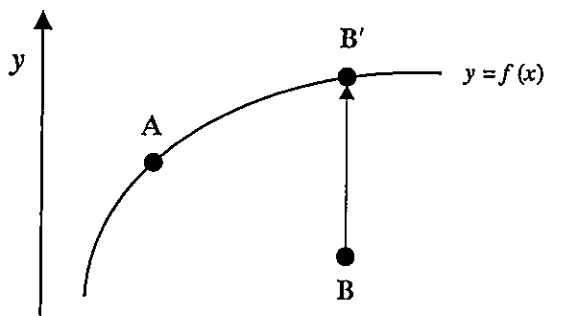

Figure 1: Scale Elasticity

Figure 1 exhibits such a sample curve $y = f(x)$ to demonstrate scale elasticity in production. Scale elasticity is well-defined at a point on the efficient portion of the input-output correspondence, e.g., the point A. For an inefficient DMU operating on point such as B, ρ is defined on its upward projected point B$'$.

In the case of multiple input-output environment, although Baumol *et al.* (1988) discussed DSE in terms of cost and output, we utilize the same terminology in our case because any differentiable cost function, whatever the number of outputs involved, and whether or not it is derived from a homogeneous production process, has a local degree of homogeneity (or, equivalently DSE) that is reciprocal of RTS parameter of a production process (p.56). We deal with the production possibility set P_{BCC} related with the multiple input-output correspondence as defined by

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

where e is a row vector with all its elements being equal to one.

Analogous to the single input-output case, we measure the scale elasticity of DMUs positioned on the efficient portion of P_{BCC}. Such projection can be realized by solving the following output-oriented BCC model (Banker *et al.*, 1984) for each DMU $(\boldsymbol{x}_o, \boldsymbol{y}_o)$ $(o = 1, \ldots, n)$.

$$\begin{array}{ll} \text{[BCC-O]} & \theta^*_{BCC} = \max\ \theta_{BCC} \\ \text{subject to} & \boldsymbol{x}_o = X\boldsymbol{\lambda} + \boldsymbol{s}^- \\ & \theta_{BCC}\boldsymbol{y}_o = Y\boldsymbol{\lambda} - \boldsymbol{s}^+ \\ & e\boldsymbol{\lambda} = 1 \\ & \boldsymbol{\lambda} \geq \mathbf{0},\ \boldsymbol{s}^- \geq \mathbf{0},\ \boldsymbol{s}^+ \geq \mathbf{0}. \end{array} \tag{3}$$

To solve [BCC-O], we employ a two-stage process: first we obtain θ^*_{BCC} by solving the program [BCC-O], and then we maximize the sum of input slacks s_i^- and output slacks s_r^+ while keeping the objective function value at θ^*_{BCC}. Symbolically, we utilize the objective function of the form

$$\max\ \theta_{BCC} + \varepsilon \left(\sum_{i=1}^{m} s_i^- + \sum_{r=1}^{s} s_r^+ \right) \tag{4}$$

for the two-stage evaluation process, where ε is a Non-Archimedean small positive number. Refer to Cooper *et al.* (1999) for the detailed explanation of this two-stage evaluation process.

Let an optimal solution of [BCC-O] be $(\theta^*_{BCC}, \boldsymbol{\lambda}^*, \boldsymbol{s}^{-*}, \boldsymbol{s}^{+*})$. Then the DMU $(\boldsymbol{x}_o, \boldsymbol{y}_o)$ is called "strongly efficient" if and only if $\theta^*_{BCC} = 1$, $\boldsymbol{s}^{-*} = \boldsymbol{0}$ and $\boldsymbol{s}^{+*} = \boldsymbol{0}$ hold for every optimal solution for [BCC-O]. Otherwise, the DMU $(\boldsymbol{x}_o, \boldsymbol{y}_o)$ can be brought into the strongly efficient status by the BCC-O projection as defined by

$$\boldsymbol{x}^*_o \leftarrow \boldsymbol{x}_o - \boldsymbol{s}^{-*} \tag{5}$$

$$\boldsymbol{y}^*_o \leftarrow \theta^*_{BCC}\boldsymbol{y}_o + \boldsymbol{s}^{+*}. \tag{6}$$

See Cooper *et al.* (1999) for the detailed explanation. Hereafter, we assume that the DMU $(\boldsymbol{x}_o, \boldsymbol{y}_o)$ is strongly efficient.

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

$$\text{[Dual]} \quad \min \boldsymbol{v}\boldsymbol{x}_o - w \tag{7}$$

$$\text{subject to} \quad -\boldsymbol{v}X + \boldsymbol{u}Y + \boldsymbol{e}w \leq \boldsymbol{0} \tag{8}$$

$$\boldsymbol{u}\boldsymbol{y}_o = 1 \tag{9}$$

$$\boldsymbol{v} \geq \boldsymbol{0},\ \boldsymbol{u} \geq \boldsymbol{0},\ w : \text{free in sign.} \tag{10}$$

Since 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

$$\boldsymbol{v}^*\boldsymbol{x}_o - w^* = 1. \tag{11}$$

From (8),(9) and (11), it holds, for the optimal $(\boldsymbol{v}^*, \boldsymbol{u}^*, w^*)$, that

$$-\boldsymbol{v}^*\boldsymbol{x}_j + \boldsymbol{u}^*\boldsymbol{y}_j + w^* \leq 0 \ \ (j = 1, \ldots, n) \tag{12}$$

$$-\boldsymbol{v}^*\boldsymbol{x}_o + \boldsymbol{u}^*\boldsymbol{y}_o + w^* = 0 \tag{13}$$

$$\boldsymbol{v}^* \geq \boldsymbol{0},\ \boldsymbol{u}^* \geq \boldsymbol{0}. \tag{14}$$

This means that 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

$$\xi = \boldsymbol{v}^*\boldsymbol{x} \ \text{ and } \ \eta = \boldsymbol{u}^*\boldsymbol{y}. \tag{15}$$

Thus, the hyperplane can be expressed as

$$-\xi + \eta + w^* = 0. \tag{16}$$

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

$$\mathrm{MP} = \frac{d\eta}{d\xi} = 1, \tag{17}$$

and average product (AP) as

$$\mathrm{AP} = \frac{\eta}{\xi} = \frac{\eta}{\eta + w^*}. \tag{18}$$

At $(\boldsymbol{x}_o, \boldsymbol{y}_o)$, we have $\eta = 1$ from (9), and hence the scale elasticity (ρ) or DSE at $(\boldsymbol{x}_o, \boldsymbol{y}_o)$ is expressed by

$$\rho = 1 + w^*. \tag{19}$$

It is to be noted here that as pointed out by Førsund and Hjalmarsson (2002), ρ does not satisfy fully the requirement of RUPL as

$$\frac{\partial \rho}{\partial x_i} = -\frac{w^* v_i^*}{(\sum_i^m v_i^* x_i - w^*)^2},\ i = 1, \ldots, m$$

IRS ($w^* > 0$) implies decreasing production elasticity in accordance with RUPL, while DRS ($w^* < 0$) implies an increasing ρ, thus violating the law.

In many occasion, the optimal w^* is not uniquely determined, i.e., there exist multiple optima. In such cases, we can find the upper (lower) bound $\bar{w}$ ($\underline{w}$) of w by solving the following linear program in $\boldsymbol{v}, \boldsymbol{u}$ and w.

$$\begin{aligned}
&[\text{Upper(Lower)}] \quad && \bar{w}(\underline{w}) = \max(\min) w && (20)\\
&\text{subject to} && -\boldsymbol{v}X + \boldsymbol{u}Y + \boldsymbol{e}w \leq \boldsymbol{0} && (21)\\
& && -\boldsymbol{v}\boldsymbol{x}_o + \boldsymbol{u}\boldsymbol{y}_o + w = 0 && (22)\\
& && \boldsymbol{u}\boldsymbol{y}_o = 1 && (23)\\
& && \boldsymbol{v} \geq \boldsymbol{0},\ \boldsymbol{u} \geq \boldsymbol{0},\ w : \text{free in sign.} && (24)
\end{aligned}$$

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

$$\bar{\rho} = 1 + \bar{w} \ \text{ and } \ \underline{\rho} = 1 + \underline{w}. \tag{25}$$

Several authors have derived this same scale elasticity formulas in different ways. Førsund (1996) has found (19) assuming unique optimal solution to hold for DMU under evaluation whereas Sueyoshi (1999) and Fukuyama (2001) devised (25) for the multiple optima case. Also is found in Fukuyama (2001) the estimation of scale elasticity in Russell and Additive DEA models. However, we need to mention here that our approach towards the derivation of scale elasticity is much simpler.

It should be noted that the status of *increasing, constant* and *decreasing* returns to scale is not "absolute" one. If we impose constraints on the dual variables $\boldsymbol{v}$ and $\boldsymbol{u}$ as we do so in the case of "Assurance Region," the status of a DMU on its DSE may suffer a change. See Tone (2001) for details.

Regarding the scale elasticity in production, we have a units invariance theorem as follows:

Theorem 1 *The scale elasticity in production (ρ) as defined by (19), and its upper (lower) bound $\bar{\rho}$ ($\underline{\rho}$) are all units invariant.*

Proof: Let us change the units of X and Y respectively by diagonal matrices $D_x \in R^{m\times m}$ and $D_y \in R^{s\times s}$ as

$$X' = D_x X \quad \text{and} \quad Y' = D_y Y.$$

By this units change, [BCC-O] turns out as

$$\begin{aligned}
\text{[BCC-O']} \quad & \max\ \theta && (26)\\
\text{subject to} \quad & \boldsymbol{x}'_o = X'\boldsymbol{\lambda} + \boldsymbol{s}^- && (27)\\
& \theta \boldsymbol{y}'_o = Y'\boldsymbol{\lambda} - \boldsymbol{s}^+ && (28)\\
& \boldsymbol{e\lambda} = 1 &&\\
& \boldsymbol{\lambda} \geq \boldsymbol{0},\ \boldsymbol{s}^- \geq \boldsymbol{0},\ \boldsymbol{s}^+ \geq \boldsymbol{0}. &&
\end{aligned}$$

By multiplying D_x^{-1} and D_y^{-1} from the left to (27) and (28) respectively, we obtain the essentially same expression as those in [BCC-O]. Thus, [BCC-O] suffers from no effect.

On the other hand, the [Dual] results in

$$\begin{aligned}
\text{[Dual']} \quad & \min \boldsymbol{v}\boldsymbol{x}'_o - w && (29)\\
\text{subject to} \quad & -\boldsymbol{v}X' + \boldsymbol{u}Y' + \boldsymbol{e}w \leq \boldsymbol{0} && (30)\\
& \boldsymbol{u}\boldsymbol{y}'_o = 1 && (31)\\
& \boldsymbol{v} \geq \boldsymbol{0},\ \boldsymbol{u} \geq \boldsymbol{0},\ w : \text{free in sign}, && (32)
\end{aligned}$$

which is equivalent to following model:

$$\min \boldsymbol{v}D_x\boldsymbol{x}_o - w \tag{33}$$

$$\text{subject to} \quad -\boldsymbol{v}D_xX + \boldsymbol{u}D_yY + \boldsymbol{e}w \leq 0 \tag{34}$$

$$\boldsymbol{u}D_y\boldsymbol{y}_o = 1 \tag{35}$$

$$\boldsymbol{v} \geq 0, \ \boldsymbol{u} \geq 0, \ w : \text{free in sign.} \tag{36}$$

Let us replace $\boldsymbol{v}D_x$ by $\boldsymbol{v}'$ and $\boldsymbol{u}D_y$ by $\boldsymbol{u}'$. Then, the dual program becomes

$$\min \boldsymbol{v}'\boldsymbol{x}_o - w \tag{37}$$

$$\text{subject to} \quad -\boldsymbol{v}'X + \boldsymbol{u}'Y + \boldsymbol{e}w \leq 0 \tag{38}$$

$$\boldsymbol{u}'\boldsymbol{y}_o = 1 \tag{39}$$

$$\boldsymbol{v}' \geq 0, \ \boldsymbol{u}' \geq 0, \ w : \text{free in sign.} \tag{40}$$

Thus, we have now the same program as that in the [Dual], which concludes the proof. □

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, since we have

$$\rho = 1 + w^* = \boldsymbol{v}^*\boldsymbol{x}_o \geq 0.$$

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. Figure 2 exhibits such a phenomenon. An increase in input x results in a decrease in output y as is shown in case of points such as F and G.

In order to potentially deal with such situation, we need to modify our production possibility set as follows:

$$P_{convex} = \{(\boldsymbol{x}, \boldsymbol{y}) | \boldsymbol{x} = X\boldsymbol{\lambda}, \boldsymbol{y} \leq Y\boldsymbol{\lambda}, \boldsymbol{e}\boldsymbol{\lambda} = 1, \boldsymbol{\lambda} \geq 0\}. \tag{41}$$

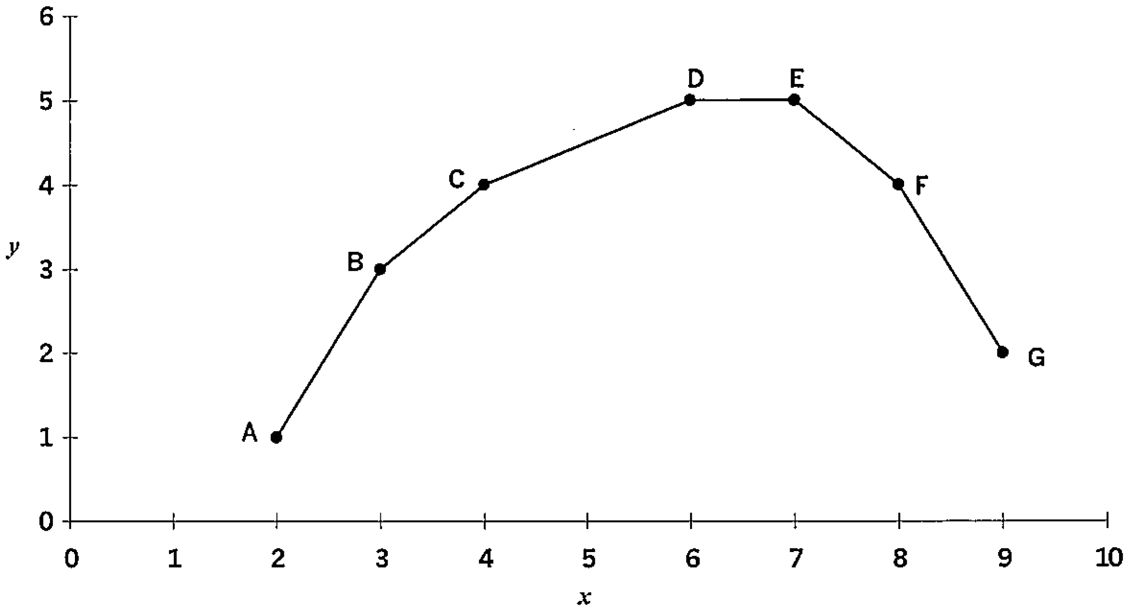

Figure 2: Congestion

In this section, 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. Similar to the P_{BCC} case, we assume here that the concerned DMU $(\boldsymbol{x}_o, \boldsymbol{y}_o)$ is on the strongly efficient frontier of P_{convex}. This means that the following model has a solution $(\theta^* = 1, \boldsymbol{s}^{+*} = \boldsymbol{0})$ for every optimal one.

$$\begin{array}{rll} \text{[Congestion-1]} & \max\ \theta & \\ \text{subject to} & \boldsymbol{x}_o = X\boldsymbol{\lambda} & \\ & \theta\boldsymbol{y}_o = Y\boldsymbol{\lambda} - \boldsymbol{s}^+ & \\ & \boldsymbol{e}\boldsymbol{\lambda} = 1 & \\ & \boldsymbol{\lambda} \geq \boldsymbol{0},\ \boldsymbol{s}^+ \geq \boldsymbol{0}. & \end{array} \tag{42}$$

If the DMU $(\boldsymbol{x}_o, \boldsymbol{y}_o)$ is not efficient, we project it onto the efficient frontier of P_{convex} by the following formulae:

$$\boldsymbol{x}_o^* \leftarrow \boldsymbol{x}_o \ \text{ unchanged} \tag{43}$$

$$\boldsymbol{y}_o^* \leftarrow \theta^*\boldsymbol{y}_o + \boldsymbol{s}^{+*}. \tag{44}$$

The input-output vector $(\boldsymbol{x}_o^*, \boldsymbol{y}_o^*)$ is now strongly efficient.

The dual program of [Congestion-1] is described as follows:

$$
\begin{aligned}
&\text{[Dual-Congestion]} && \min \boldsymbol{v}\boldsymbol{x}_o - w && (45)\\
&\text{subject to} && -\boldsymbol{v}X + \boldsymbol{u}Y + \boldsymbol{e}w \leq \mathbf{0} && (46)\\
& && \boldsymbol{u}\boldsymbol{y}_o = 1 && (47)\\
& && \boldsymbol{u} \geq \mathbf{0},\ \boldsymbol{v},\ w : \text{free in sign.} && (48)
\end{aligned}
$$

We notice that, in this dual program $\boldsymbol{v}$ is free in sign, whereas it was constrained to be nonnegative in [Dual]. Let an optimal solution vector of [Dual-Congestion] be $(\boldsymbol{v}^*, \boldsymbol{u}^*, w^*)$. Then, we can demonstrate in the similar vein as is in the case of P_{BCC} that the hyperplane $-\boldsymbol{v}^*\boldsymbol{x} + \boldsymbol{u}^*\boldsymbol{y} + w^* = 0$ is a supporting hyperplane to P_{convex} at $(\boldsymbol{x}_o, \boldsymbol{y}_o)$. The production elasticity can then be obtained by

$$\rho = 1 + w^*. \tag{49}$$

The multiple optima case can be dealt with by the following program:

$$
\begin{aligned}
&\text{[Upper-w]} && \bar{w} = \max w && (50)\\
&\text{subject to} && -\boldsymbol{v}X + \boldsymbol{u}Y + \boldsymbol{e}w \leq \mathbf{0} && (51)\\
& && -\boldsymbol{v}\boldsymbol{x}_o + \boldsymbol{u}\boldsymbol{y}_o + w = 0 && (52)\\
& && \boldsymbol{u}\boldsymbol{y}_o = 1 && (53)\\
& && \boldsymbol{u} \geq \mathbf{0},\ \boldsymbol{v}, w : \text{free in sign.} && (54)
\end{aligned}
$$

Here we have an interest in the upper bound values of w^*, since this value is closely related with "congestion." The upper bound of scale elasticity is calculated by

$$\bar{\rho} = 1 + \bar{w}. \tag{55}$$

3.1 The case with a negative upper bound $\bar{\rho}$

The dual program to [Dual-Upper-w] can be described as follows:

$$
\begin{aligned}
&\text{[Dual-Upper-w]} && \min \theta_1 - \theta_2 && (56)\\
&\text{subject to} && X\boldsymbol{\lambda} = \theta_1 \boldsymbol{x}_o && (57)\\
& && Y\boldsymbol{\lambda} \geq \theta_2 \boldsymbol{y}_o && (58)\\
& && \boldsymbol{e}\boldsymbol{\lambda} = 1 + \theta_1 && (59)\\
& && \boldsymbol{\lambda} \geq \mathbf{0}. && (60)
\end{aligned}
$$

Let an optimal solution vector of [Dual-Upper-w] be $(\theta_1^*, \theta_2^*, \boldsymbol{\lambda}^*)$. Then, the upper bound of scale elasticity $\bar{\rho}$ satisfies the following:

$$\bar{\rho} = 1 + \theta_1^* - \theta_2^*. \tag{61}$$

Now, suppose that $\bar{\rho} < 0$, i.e., $1 + \theta_1^* - \theta_2^* < 0$. In this case, from the constraints in [Dual-Upper-w] and the assumptions $X > O$ and $Y > O$, we have:

$$\theta_2^* > 1 + \theta_1^* > \theta_1^* > 0 \text{ and } \boldsymbol{\lambda}^* \neq \mathbf{0}. \tag{62}$$

Let us define $\widehat{\boldsymbol{\lambda}}$, $\widehat{\boldsymbol{x}}_o$ and $\widehat{\boldsymbol{y}}_o$ respectively as

$$\widehat{\boldsymbol{\lambda}} = \frac{1}{1+\theta_1^*}\boldsymbol{\lambda}^*, \ \widehat{\boldsymbol{x}}_o = X\widehat{\boldsymbol{\lambda}} \text{ and } \widehat{\boldsymbol{y}}_o = Y\widehat{\boldsymbol{\lambda}}. \tag{63}$$

Then, we have

$$\boldsymbol{e}\widehat{\boldsymbol{\lambda}} = 1 \tag{64}$$

$$X\widehat{\boldsymbol{\lambda}} = \widehat{\boldsymbol{x}}_o = \frac{\theta_1^*}{1+\theta_1^*}\boldsymbol{x}_o < \boldsymbol{x}_o \tag{65}$$

$$Y\widehat{\boldsymbol{\lambda}} = \widehat{\boldsymbol{y}}_o \geq \frac{\theta_2^*}{1+\theta_1^*}\boldsymbol{y}_o > \boldsymbol{y}_o. \tag{66}$$

Thus, there exists $(\widehat{\boldsymbol{x}}_o, \widehat{\boldsymbol{y}}_o) \in P_{convex}$ such that $\widehat{\boldsymbol{x}}_o < \boldsymbol{x}_o$ and $\widehat{\boldsymbol{y}}_o > \boldsymbol{y}_o$. This means that the DMU $(\boldsymbol{x}_o, \boldsymbol{y}_o)$ is in the region of "congestion" where there exists a different activity $(\widehat{\boldsymbol{x}}_o, \widehat{\boldsymbol{y}}_o)$ that uses less inputs to produce more outputs.

Conversely, suppose that there exists $(\tilde{\boldsymbol{x}}_o, \tilde{\boldsymbol{y}}_o) \in P_{convex}$ such that

$$\tilde{\boldsymbol{x}}_o = \alpha\boldsymbol{x}_o \text{ (with } 0 < \alpha < 1) \text{ and } \tilde{\boldsymbol{y}}_o \geq \beta\boldsymbol{y}_o \text{ (with } \beta > 1). \tag{67}$$

Then, we can prove that $(\boldsymbol{x}_o, \boldsymbol{y}_o)$ has a negative upper bound of scale elasticity, as follows:

Since adding the activity $(\tilde{\boldsymbol{x}}_o, \tilde{\boldsymbol{y}}_o)$ to the data set (X, Y) has no effect on the curvature of the set P_{convex}, we enlarge the data set (X, Y) to $(\bar{X}, \bar{Y})$ by adding ($\tilde{\boldsymbol{x}}_o$ and $\tilde{\boldsymbol{y}}_o$) to the top of X and Y, respectively. Thus, we have

$$\bar{X} = (\tilde{\boldsymbol{x}}_o, \boldsymbol{x}_1, \ldots, \boldsymbol{x}_n) \in R^{m\times(n+1)} \text{ and } \bar{Y} = (\tilde{\boldsymbol{y}}_o, \boldsymbol{y}_1, \ldots, \boldsymbol{y}_n) \in R^{s\times(n+1)}.$$

Let us define the corresponding $\bar{\boldsymbol{\lambda}} \in R^{n+1}$ by

$$\bar{\boldsymbol{\lambda}}^T = (\lambda_1, 0, 0, \ldots, 0).$$

Then, we have

$$\bar{X}\bar{\boldsymbol{\lambda}} = \lambda_1 \tilde{\boldsymbol{x}}_o = \lambda_1 \alpha \boldsymbol{x}_o \tag{68}$$
$$\bar{Y}\bar{\boldsymbol{\lambda}} = \lambda_1 \tilde{\boldsymbol{y}}_o \geq \lambda_1 \beta \boldsymbol{y}_o \tag{69}$$
$$\boldsymbol{e}\bar{\boldsymbol{\lambda}} = \lambda_1. \tag{70}$$

Let us define λ_1, θ_1 and θ_2 respectively by

$$\lambda_1 = \frac{1}{1-\alpha} \tag{71}$$
$$\theta_1 = \alpha\lambda_1 = \frac{\alpha}{1-\alpha} \tag{72}$$
$$\theta_2 = \beta\lambda_1 = \frac{\beta}{1-\alpha}. \tag{73}$$

Then, we have

$$\bar{X}\bar{\boldsymbol{\lambda}} = \theta_1 \boldsymbol{x}_o \tag{74}$$
$$\bar{Y}\bar{\boldsymbol{\lambda}} \geq \theta_2 \boldsymbol{y}_o \tag{75}$$
$$\boldsymbol{e}\bar{\boldsymbol{\lambda}} = 1 + \theta_1 \tag{76}$$
$$\bar{\boldsymbol{\lambda}} \geq \boldsymbol{0}. \tag{77}$$

Thus, this solution vector $(\theta_1, \theta_2, \bar{\boldsymbol{\lambda}})$ is feasible in [Dual-Upper-w]. For this feasible solution we now have

$$1 + \theta_1 - \theta_2 = \frac{1-\beta}{1-\alpha} < 0, \text{ since } \alpha < 1 \text{ and } \beta > 1 \text{ by assumption.} \tag{78}$$

Hence, it is proved that the DMU $(\boldsymbol{x}_o, \boldsymbol{y}_o)$ has a negative upper bound of scale elasticity.

To conclude this section, we have the following definition and theorem.

Definition 1 *(Strong Congestion) A DMU $(\boldsymbol{x}_o, \boldsymbol{y}_o)$ is strongly congested if there exists an activity $(\tilde{\boldsymbol{x}}_o, \tilde{\boldsymbol{y}}_o) \in P_{convex}$ such that $\tilde{\boldsymbol{x}}_o = \alpha \boldsymbol{x}_o$ (with $0 < \alpha < 1$) and $\tilde{\boldsymbol{y}}_o \geq \beta \boldsymbol{y}_o$ (with $\beta > 1$). at $(\boldsymbol{x}_o, \boldsymbol{y}_o)$*

Theorem 2 *A DMU $(\boldsymbol{x}_o, \boldsymbol{y}_o)$ is in the status of strong congestion, if and only if the upper scale elasticity $(\bar{\rho})$ as measured by the scheme* [Upper-w] *is negative.*

We notice that this definition of strong congestion requires existence of a DMU $(\tilde{\boldsymbol{x}}_o, \tilde{\boldsymbol{y}}_o) \in P_{convex}$ with a proportionally reduced input vector $\tilde{\boldsymbol{x}}_o$ $(= \alpha \boldsymbol{x}_o,\ 0 < \alpha < 1)$ and an enlarged output vector $\tilde{\boldsymbol{y}}_o$ $(\geq \beta \boldsymbol{y}_o,\ \beta > 1)$.

3.2 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). From this point of view, the above definition of strong congestion is too restrictive in that a proportionate reduction in *all* inputs warrants an increase in *all* outputs. We now redefine (weak) congestion by relaxing such stringent requirements in the following definition.

Definition 2 *(Weak congestion) A DMU is (weakly) congested if it is strongly efficient with respect to P_{convex} and there exist an activity in P_{convex} that uses less resources in one or more inputs for making more products in one or more outputs.*

In order to deal with this less restrictive definition, we present here an another scheme.

First, we assume that the concerned DMU $(\boldsymbol{x}_o, \boldsymbol{y}_o)$ is strongly efficient in P_{convex} as defined in (41). Thus, the program [Congestion-1] has $(\theta^* = 1, \boldsymbol{s}^{+*} = \boldsymbol{0})$ for every optimal solution. For this DMU $(\boldsymbol{x}_o, \boldsymbol{y}_o)$, we solve the following linear program with variables $\boldsymbol{\lambda}, \boldsymbol{t}^-$ and $\boldsymbol{t}^+$.

$$
\begin{aligned}
\text{[Congestion-2]} \quad & \max \ \frac{1}{s}\sum_{r=1}^{s}\frac{t_r^+}{y_{ro}} && (79)\\
\text{subject to} \quad & \boldsymbol{x}_o = X\boldsymbol{\lambda} + \boldsymbol{t}^- \\
& \boldsymbol{y}_o = Y\boldsymbol{\lambda} - \boldsymbol{t}^+ \\
& \boldsymbol{e}\boldsymbol{\lambda} = 1 \\
& \boldsymbol{\lambda} \geq \boldsymbol{0},\ \boldsymbol{t}^- \geq \boldsymbol{0},\ \boldsymbol{t}^+ \geq \boldsymbol{0}.
\end{aligned}
$$

To solve [Congestion-2], we employ a two-stage process similar to (4): first, we maximize the objective function in (79), and then we maximize $\sum_{i=1}^{m} t_i^-/x_{io}$, while keeping the objective value of (79) at the optimum level. Symbolically, we utilize the objective function of form

$$
\max \ \frac{1}{s}\sum_{r=1}^{s}\frac{t_r^+}{y_{ro}} + \varepsilon\frac{1}{m}\sum_{i=1}^{m}\frac{t_i^-}{x_{io}} \tag{80}
$$

for the two-stage evaluation process, where ε is a Non-Archimedean small positive number.

Let an optimal solution vector be $(\boldsymbol{\lambda}^*, \boldsymbol{t}^{-*}, \boldsymbol{t}^{+*})$. We have now two cases:

Case 1. $t^{+*} = 0$.

In this case, no congestion is observed in activity (x_o, y_o), since a decrease in inputs cannot increase any outputs.

Case 2. $t^{+*} \neq 0$.

In this case, t^{-*} is not even zero too, since activity (x_o, y_o) is strongly efficient in P_{convex}. Thus, we identify congestion in activity (x_o, y_o).

The objective function form in [Congestion-2] is an output-oriented version of the Slacks-based Measure (SBM) that was introduced in Tone (2001). It is units invariant.

Henceforth, we deal with the Case 2, i.e., $t^{+*} \neq 0,\ t^{-*} \neq 0$. Based on the optimal solution vector $(\lambda^*, t^{-*}, t^{+*})$, we define $\widehat{x}_o$ and $\widehat{y}_o$ as

$$\widehat{x}_o = X\lambda = x_o - t^{-*} \tag{81}$$

$$\widehat{y}_o = Y\lambda = y_o + t^{+*}. \tag{82}$$

$(\widehat{x}_o, \widehat{y}_o)$ is an improved (less congested) activity than (x_o, y_o).

As a proxy measure for scale elasticity, we propose the following formula. First, we define an approximation to the marginal production rate (MPR) as

$$\text{MPR} = -\frac{1}{\bar{s}} \sum_{r=1}^{s} \frac{t_r^{+*}}{y_{ro}} \bigg/ \frac{1}{\bar{m}} \sum_{i=1}^{m} \frac{t_i^{-*}}{x_{io}}, \tag{83}$$

where $\bar{s}$ and $\bar{m}$ are the numbers of positive t_r^{+*} $(r = 1, \ldots, s)$ and positive t_i^{-*} $(i = 1, \ldots, m)$, respectively. The average production rate (APR) is defined as

$$\text{APR} = \frac{1}{s} \sum_{r=1}^{s} \frac{y_{ro}}{y_{ro}} \bigg/ \frac{1}{m} \sum_{i=1}^{m} \frac{x_{io}}{x_{io}} = 1. \tag{84}$$

Thus, we have the following approximation measure for DSE, which is given below.

$$\text{DSE} = \frac{\text{MPR}}{\text{APR}} = -\frac{1}{\bar{s}} \sum_{r=1}^{s} \frac{t_r^{+*}}{y_{ro}} \bigg/ \frac{1}{\bar{m}} \sum_{i=1}^{m} \frac{t_i^{-*}}{x_{io}}. \tag{85}$$

This can be interpreted as the ratio of the average improvement in outputs to the average reduction in inputs. To note here that the negaive DSE value for any activity indicates that congestion is present in that activity.

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

Theorem 3 *Suppose that the DMU (x_o, 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 $s^{+*} \neq \mathbf{0}$) by the model* [BCC-O] *in (3).*

Proof : The status of weak congestion has two alternatives: (A) Congested and (B) Not congested, while the optimal solution for the [BCC-O] model has three alternatives: (a) $\theta^*_{BCC} > 1$, (b) $\theta^*_{BCC} = 1$ and $s^{+*} \neq \mathbf{0}$, and (c) $\theta^*_{BCC} = 1$ and $s^{+*} = \mathbf{0}$. We demonstrate that (A) corresponds to (a) or (b), and (B) corresponds to (c).

In case of (a), the second constraint in (3) can be transformed into:

$$y_o = X\lambda^* - (\theta^*_{BCC} - 1)y_o - s^{+*}.$$

Since $\theta^*_{BCC} > 1$, it holds that $(\theta^*_{BCC} - 1)y_o + s^{+*} > \mathbf{0}$ and hence the weak congestion model [Congestion-2] has a positive objective function value, resulting in (A). In case of (b), the DMU (x_o, y_o) is weakly congested by definition and hence (b) corresponds to (A). In case of (c), no congestion is identified, and (c) corresponds to (B). This concludes the proof. □

Regarding the improved DMU $(\hat{x}_o, \hat{y}_o)$, we have the following theorem:

Theorem 4 *The improved DMU $(\hat{x}_o, \hat{y}_o)$ defined by (81) and (82) is not weakly congested.*

Proof : The congestion status of the $(\hat{x}_o, \hat{y}_o)$ is evaluated by the program below:

$$\begin{aligned} & \max \quad \frac{1}{s}\sum_{r=1}^{s}\frac{\tau_r^+}{y_{ro}} && (86) \\ \text{subject to} \quad & \hat{x}_o = X\mu + \tau^- \\ & \hat{y}_o = Y\mu - \tau^+ \\ & e\mu = 1 \\ & \mu \geq \mathbf{0},\ \tau^- \geq \mathbf{0},\ \tau^+ \geq \mathbf{0}. \end{aligned}$$

If it has an optimal solution $(\mu^*, \tau^{-*}, \tau^{+*})$ with $\tau^{+*} \neq \mathbf{0}$, then, referring to (81) and (82), it holds that

$$\begin{aligned} x_o &= X\mu^* + t^{-*} + \tau^{-*} \\ y_o &= Y\mu^* - t^{+*} - \tau^{+*}. \end{aligned}$$

Since $t^{+*} + \tau^{+*} \geq t^{+*}$ and $t^{+*} + \tau^{+*} \neq t^{+*}$, this contradicts the maximality of t^{+*} for [Congestion-2]. Thus, we have $\tau^{+*} = \mathbf{0}$ and hence $(\widehat{x}_o, \widehat{y}_o)$ is not weakly congested. □

If an increase in some special outputs is required, the weights on the respective output slacks t_r^+ in the objective function (79) can be assigned in the following manner:

$$\max \frac{1}{s} \sum_{r=1}^{s} \frac{w_r t_r^+}{y_{ro}},$$

where the weight $w_r \geq 0$ (with $\sum_{r=1}^{s} w_r = s$) is the weight assigned to the output r.

4 An Illustrative Example

We now illustrate the above procedure with the help of input-output data of seven DMUs that were used in Figure 2. Table 1 reports the DSE results obtained from [BCC-O], [Congestion-1] (Strong) and [Congestion-2] (Weak) models. Depending on $\bar{\rho}$ and $\underline{\rho}$ in [BCC-O] model, we identified DMU A

Table 1: An Illustrative Example

	Data		[BCC-O]			[Congestion-1]			[Congestion-2]	
DMU	Input	Output	$\bar{\rho}$	$\underline{\rho}$	RTS	$\bar{\rho}$	$\underline{\rho}$	RTS	ρ	RST
A	2	1	∞	4	IRS	∞	4	IRS	-	IRS
B	3	3	2	1	CRS	2	1	CRS	1.5	CRS
C	4	4	1	0.5	CRS	1	0.5	CRS	0.75	CRS
D	6	5	0.6	0	DRS	0.6	0	DRS	0.3	DRS
E	7	5	0.6	0	DRS	0	-1.4	DRS	-0.7	DRS
F	8	4	0.6	0	DRS	-2	-4	Congestion	-1	Congestion
G	9	2	0.6	0	DRS	-9	$-\infty$	Congestion	-4.5	Congestion

operating under IRS, DMUs B and C under CRS, and the remaining DMUs under DRS. In this model, however, DMUs E, F and G are inefficient, i.e., they are not on the strongly efficient portion of P_{BCC}, and hence we projected them to the frontier point (DMU D) using (5) and (6). Thus, DMUs E, F and G share same information with DMU D with respect to $\bar{\rho}$, $\underline{\rho}$ and RTS. However, in [Congestion-1] and [Congestion-2] models, all the DMUs are strongly efficient with respect to P_{convex}.

[Congestion-1] evaluated the upper elasticity $\bar{\rho}$ via (50)-(55), and the lower elasticity $\underline{\rho}$ using a similar scheme (by replacing max with min). In this model, DMUs A, B, C and D have eventually the same $\bar{\rho}$ and $\underline{\rho}$ with those in [BCC-O] model, and hence their RTS behavior in [BCC-O] and [Congestion-1] are same. DMU E has $\bar{\rho} = 0$ and $\underline{\rho} = -1.4$. We judged E being not congested, and operate under DRS. DMUs F and G are found strongly congested, because they have a negative $\bar{\rho}$.

In [Congestion-2] model, however, two DMUs F and G as found weakly congested, as is evident from their negative scale elasticity values. These values are less than the $\bar{\rho}$ values in [Congestion-1] model. This is due to the fact that the latter $\bar{\rho}$ reflects the slope of the supporting hyperplane to P_{convex} at the concerned DMU (x_o, y_o), while the former (elasticity in the weak congestion model) is calculated along the line segment connecting (x_o, y_o) and $(\widehat{x}_o, \widehat{y}_o)$ that may pass through inside P_{convex}. Other DMUs are identified as being not congested in this model, and they have the same elasticity and RTS with those in [Congestion-1]. So we recorded the average of $\bar{\rho}$ and $\underline{\rho}$ in the [Congestion-1] model as the scale elasticity value in the [Congestion-2] model.

For the congested DMUs F and G, we found in [Congestion-1] the improved DMU $(\widehat{x}_o, \widehat{y}_o)$ as designated by (63). Table 2 reports these results. As can be seen, DMU F is projected to DMU E (7, 5), and DMU G to DMU F (8, 4), which is still congested.

Table 2: Improvement of Congested DMUs

	Data		[Congestion-1] Improved		[Congestion-2] Improved	
DMU	Input	Output	Input	Output	Input	Output
F	8	4	7	5	6	5
G	9	2	8	4	6	5

In the weak congestion case ([Congestion-2]), however, the improved input and output combination using the formulas (81) and (82) is the activity (6, 5) that is DMU D, which is not weakly congested, as asserted by Theorem 4. See Table 2.

5 An Empirical Study

We analyze here the scale elasticity change over time of the operations of chain stores (i.e., supermarkets) in Japan for a period of 27 years from 1975 through 2001. For the computation of scale elasticity, we have considered here one output: annual sales (unit: hundred million yen), and two inputs: the number of stores and the total area of stores (unit: 1000m^2).

Table 3: Chain Stores Data Set

Year	Number	Area	Sales	Sales/Number	Sales/Area
75	2,412	5,480	41,091	17.036	7.498
76	3,163	6,233	48,367	15.291	7.760
77	3,350	6,798	56,000	16.716	8.238
78	3,371	7,274	60,940	18.078	8.378
79	3,778	7,992	69,046	18.276	8.639
80	4,020	8,500	77,347	19.241	9.100
81	5,029	9,246	85,805	17.062	9.280
82	5,164	9,639	90,433	17.512	9.382
83	5,285	9,981	95,640	18.096	9.582
84	5,618	10,276	100,257	17.846	9.756
85	5,981	10,521	105,944	17.713	10.070
86	6,217	10,766	109,857	17.670	10.204
87	6,455	11,144	116,114	17.988	10.419
88	6,674	11,418	125,404	18.790	10.983
89	6,829	11,717	131,862	19.309	11.254
90	6,995	11,987	140,817	20.131	11.747
91	7,338	12,463	150,583	20.521	12.082
92	7,946	13,426	152,943	19.248	11.392
93	8,236	14,147	155,128	18.835	10.965
94	7,722	15,014	158,714	20.553	10.571
95	7,727	15,022	161,739	20.932	10.767
96	7,822	16,191	169,786	21.706	10.486
97	7,531	16,969	167,195	22.201	9.853
98	7,201	17,627	167,187	23.217	9.485
99	7,281	18,364	165,480	22.728	9.011
00	7,053	19,698	162,847	23.089	8.267
01	6,067	16,176	154,671	25.494	9.562
Average	6,010	12,003	119,304	19.455	9.805

Table 3 report such data. The data were directly obtained from the Japan

Chain Stores Association. This Association covers respectively roughly about 20%, 42% and 51% of total Number, Area and Sales of supermarkets as of 1999.

Let us first analyze the behavior of the trends of the inputs and output. As is seen in Figure 3, we find there is a slow but steady rise in the number of chain stores till 1993 after which the trend continues declining consistently. Barring last year of our study, though the total area is seen to be consistently rising throughout, the first-half of our total sample period is characterized by logarithmic rise whereas the second-half experiences an exponential rise. This indicates the rapid growth of large-scale stores as contrasted to remarkable extinction of small stores in recent years. As regards the nature of sales trend, we find that it has shown an increasing trend until 1996 after which the trend is seen to be consistently declining. This reflects the "bear" consumer minds and weak Japan economies in recent years.

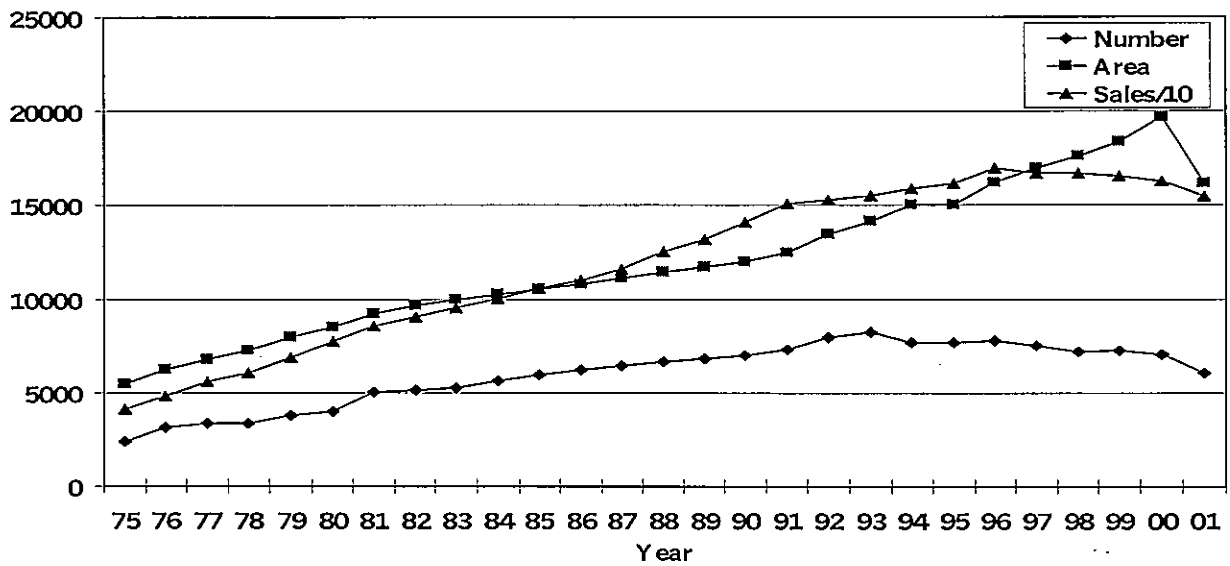

Figure 3: Trend

In Figure 4 we see that while the trend for sales per store is of rise throughout, this is not so in case of sales per area, which has exhibited an upward trend until 1991 after which it shows a declining trend. The year 2001 however, demonstrates an improvement in both indices.

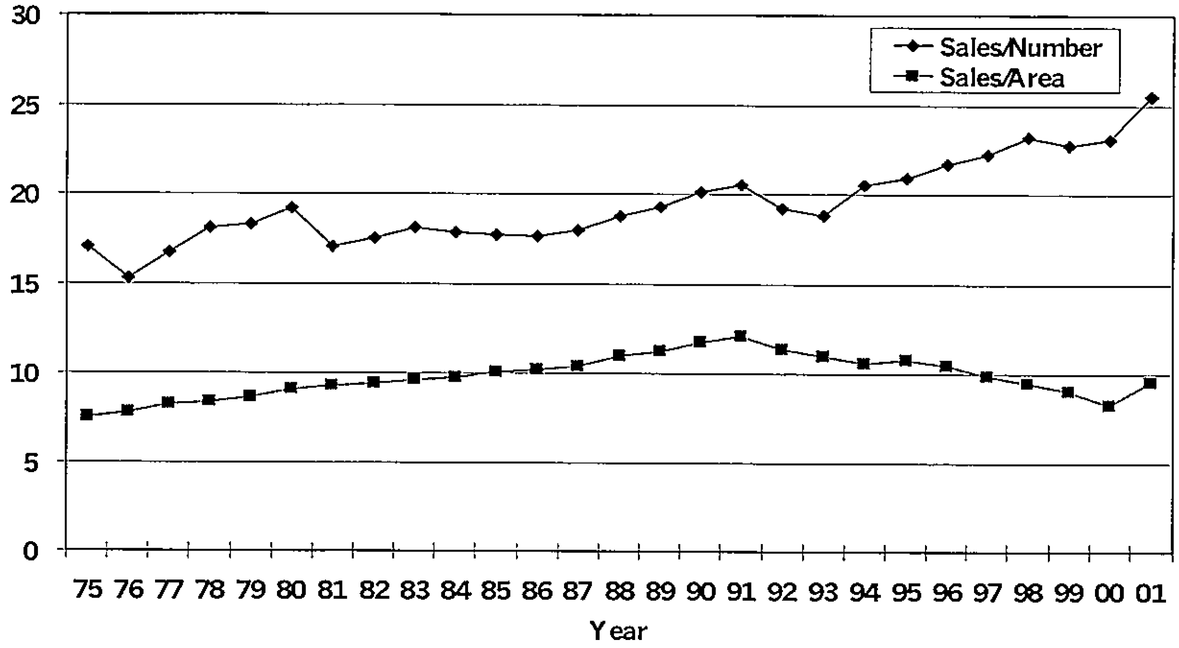

Figure 4: Sales/Number and Sales/Area

Based on this data set, we evaluated the scale elasticity of operations over time, with each year being treated as distinct DMU. Three DEA models were utilized for this purpose: the BCC-O, the Strong and Weak Congestion Models.

5.1 Results of the BCC-O model

We applied the [BCC-O] model (3) to the data set to compute the efficiency score, input excesses and output shortfalls, based on which the projected inputs and outputs are computed using the formulas (5) and (6). We then solved the [Upper(Lower)] program (20)-(24) to compute the upper and lower bound of scale elasticity $\bar{\rho}$ ($\underline{\rho}$) using (25). Finally, the average elasticity score is computed as the mean the lower and upper scale elasticity values. However, in case the upper bound is ∞, the lower bound is taken as the average. These results are reported in Table 4. The average BCC-O scores is found to be 1.06, which indicates that the chain stores have been operating their business in a considerably efficient way throughout 27 years. Concerning the scale elasticity of operations of chain stores, we find that barring for the last fours years of our study period it has overall shown a declining trend. However, the chain stores were operating under IRS for the first 16 years (1975-1990) followed by CRS in 1991 after which DRS sets in. The year 1990-1991, which corresponds to the collapse of the bubble economies in Japan is the turning point. The scale elasticity score, which was continuously falling since the begining period of our study, started showing increasing trend particularly after 1998. This finding can be viewed as the importance level of chain stores' scale expansion reflecting a signing of recovery resulting from intense business restructuring in recent years. The BCC-O analysis suggests that the Japan Chain Stores Association, which was enjoying scale economies until 1990, started suffering from the scale diseconomies after 1992. Rapid expansions in the number as well as size of chain stores were not accompanied with the corresponding growth in sales. As a result, several supermarkets were forced to go bankrupt. This also suggests existence of congestion in this business.

Table 4: BCC-O Results

DMU	Score	Projected Number	Projected Area	Projected Sales	Elasticity Upper($\bar{\rho}$)	Elasticity Lower($\underline{\rho}$)	Elasticity Average	RTS
75	1.000	2,412	5,480	41,091	∞	1.599	1.599	IRS
76	1.094	2,943	6,233	52,898	1.848	1.465	1.656	IRS
77	1.103	3,342	6,798	61,757	1.726	1.398	1.562	IRS
78	1.066	3,371	7,274	64,955	1.379	1.379	1.379	IRS
79	1.084	3,778	7,992	74,829	1.329	1.329	1.329	IRS
80	1.049	4,020	8,500	81,175	1.303	1.303	1.303	IRS
81	1.161	5,029	9,246	99,590	1.247	1.247	1.247	IRS
82	1.148	5,164	9,639	103,773	1.237	1.237	1.237	IRS
83	1.124	5,285	9,981	107,462	1.229	1.229	1.229	IRS
84	1.135	5,618	10,276	113,825	1.216	1.216	1.216	IRS
85	1.134	5,968	10,521	120,133	1.373	1.205	1.289	IRS
86	1.129	6,141	10,766	123,974	1.362	1.198	1.280	IRS
87	1.119	6,408	11,144	129,901	1.345	1.189	1.267	IRS
88	1.070	6,601	11,418	134,198	1.334	1.183	1.259	IRS
89	1.053	6,812	11,717	138,886	1.323	1.177	1.250	IRS
90	1.016	6,995	11,987	143,019	1.172	1.172	1.172	IRS
91	1.000	7,338	12,463	150,583	1.298	0.426	0.862	CRS
92	1.017	7,463	13,426	155,543	0.760	0.445	0.602	DRS
93	1.027	7,557	14,147	159,257	0.766	0.458	0.612	DRS
94	1.032	7,669	15,014	163,723	0.772	0.472	0.622	DRS
95	1.013	7,670	15,022	163,764	0.772	0.472	0.622	DRS
96	1.000	7,822	16,191	169,786	0.780	0.000	0.390	DRS
97	1.008	7,531	16,864	168,568	0.575	0.187	0.381	DRS
98	1.000	7,201	17,627	167,187	0.571	0.180	0.376	DRS
99	1.012	7,281	17,442	167,522	0.572	0.182	0.377	DRS
00	1.017	7,053	17,438	165,554	0.567	0.470	0.519	DRS
01	1.000	6,067	16,176	154,671	1.219	0.433	0.826	CRS
Average	1.060							

5.2 Results of the strong congestion model

In order to investigate the existence of strong congestion as defined in Section 3.1, we analyzed the data set in Table 3 with the convex production possibility set assumption (41). The results are reported in Table 5. First, we solved the [Congestion-1] (42) to compute the technical efficiency θ^*, which is shown in the column 'Score' of this table. As expected, this value is not greater than that in the [BCC-O] model. The 'Projected Sales' is calculated using the formula (44), while Number and Area remain unchanged. Using these projected data, we calculated the upper scale elasticity by means of the program [Upper-w] in (50)-(54) and (55), and lower elasticity using similar scheme. These scores are reported in the columns 'Elasticity Upper' and 'Elasticity Lower'. Theorem 2 asserts that strong congestion occurs if and only if $\bar{\rho}$ is negative. We found two years 1997 ($\bar{\rho} = -0.252$) and 1999 ($\bar{\rho} = -0.274$) to be strongly congested.

The improved activities, using (65) and (66), for these two years are:

For 1997:

$$\text{Number} = 75,07(-0.3\%), \ \text{Area} = 16,917(-0.3\%), \ \text{Sales} = 168,472(0.07\%)$$

For 1999:

$$\text{Number} = 71,69(-1.5\%), \ \text{Area} = 18,080(-1.5\%), \ \text{Sales} = 166,237(0.42\%).$$

Numbers in parentheses indicate % change. As defined in Section 3.1, the strong congestion is identified if and only if a reduction in *all* inputs requires an increase in *all* outputs. This restrictive definition seems to limit the numbers of congested DMUs to a small number, in which case % changes are small too.

5.3 Results of the weak congestion model

This model identifies congestion when a reduction in *some* inputs causes an increase in *some* outputs, thus resulting in an occurrence of more congested DMUs than in the case of strong congestion. Table 6 exhibits the results where 14 years of operation are found to be under congestion. The column 'Slacks Sales' indicates shortfalls in the output (Sales), while 'Slacks Number' and 'Slacks Area' correspond to excesses in the inputs (Number of stores and

Table 5: Strong Congestion Results

DMU	Score	Projected Number	Projected Area	Projected Sales	Elasticity Upper	Elasticity Lower	Elasticity Average	RTS
75	1.000	2,412	5,480	41,091	∞	1.599	1.599	IRS
76	1.000	3,163	6,233	48,367	∞	2.546	2.546	IRS
77	1.100	3,350	6,798	61,587	2.215	2.215	2.215	IRS
78	1.066	3,371	7,274	64,955	1.379	1.379	1.379	IRS
79	1.084	3,778	7,992	74,829	1.329	1.329	1.329	IRS
80	1.049	4,020	8,500	81,175	1.303	1.303	1.303	IRS
81	1.161	5,029	9,246	99,590	1.247	1.247	1.247	IRS
82	1.148	5,164	9,639	103,773	1.237	1.237	1.237	IRS
83	1.124	5,285	9,981	107,462	1.229	1.229	1.229	IRS
84	1.135	5,618	10,276	113,825	1.216	1.216	1.216	IRS
85	1.131	5,981	10,521	119,866	1.624	1.624	1.624	IRS
86	1.085	6,217	10,766	119,147	3.336	3.336	3.336	IRS
87	1.109	6,455	11,144	128,738	3.162	3.162	3.162	IRS
88	1.000	6,674	11,418	125,404	∞	3.219	3.219	IRS
89	1.051	6,829	11,717	138,530	1.540	1.540	1.540	IRS
90	1.016	6,995	11,987	143,019	1.172	1.172	1.172	IRS
91	1.000	7,338	12,463	150,583	∞	0.222	0.222	CRS
92	1.000	7,946	13,426	152,943	∞	$-\infty$	1.000	CRS
93	1.000	8,236	14,147	155,128	0.164	$-\infty$	0.164	DRS
94	1.030	7,722	15,014	163,439	0.283	0.283	0.283	DRS
95	1.011	7,727	15,022	163,459	0.283	0.283	0.283	. DRS
96	1.000	7,822	16,191	169,786	0.780	$-\infty$	0.780	DRS
97	1.007	7,531	16,969	168,342	-0.252	-0.252	-0.252	Congestion
98	1.000	7,201	17,627	167,187	0.571	-0.261	0.155	DRS
99	1.000	7,281	18,364	165,538	-0.274	-0.274	-0.274	Congestion
00	1.000	7,053	19,698	162,847	∞	$-\infty$	1.000	CRS
01	1.000	6,067	16,176	154,671	1.579	0.368	0.973	CRS
Average	1.048							

Area of stores), which are all displayed with negative values. The 'Improved' inputs and output are computed using formulas (81) and (82), where as 'Elasticity' is by (85). It is interesting to note that, before 1995 congestion was due to the excess in number of stores, whereas after 1997 it is due to the excess in area of stores. This finding tallies with the trend of both the inputs in Figure 4.

5.4 Summary of this case study

We conclude this section by comparing three models with respect to their average scale elasticity score trends in Figure 5.

1. The scale elasticity scores in [BCC-O] model indicate that Supermarkets operate under IRS upto 1990, CRS in 1991 and DRS for the remaining years barring the last year of operation. The managerial implication is that the scale expansion after 1990 was not rewarding when several supermarkets were obliged to go bankrupt in 90s.

2. The strong congestion model identified two years of chain stores' operation in the late 90s as congested, reflecting excesses in both the inputs

Table 6: Weak Congestion Results

DMU	Score	Slacks Number	Slacks Area	Slacks Sales	Improved Number	Improved Area	Improved Sales	Elasticity	Congestion
75	1.000	0	0	0	2,412	5,480	41,091		
76	1.000	-219.81	0	4530.88	2,943	6,233	52,898	-1.348	Congestion
77	1.100	-8.25	0	169.98	3,342	6,798	61,757	-1.121	Congestion
78	1.066	0	0	0	3,371	7,274	64,955		
79	1.084	0	0	0	3,778	7,992	74,829		
80	1.049	0	0	0	4,020	8,500	81,175		
81	1.161	0	0	0	5,029	9,246	99,590		
82	1.148	0	0	0	5,164	9,639	103,773		
83	1.124	0	0	0	5,285	9,981	107,462		
84	1.135	0	0	0	5,618	10,276	113,825		
85	1.131	-12.94	0	266.73	5,968	10,521	120,133	-1.029	Congestion
86	1.085	-76.11	0	4827.11	6,141	10,766	123,974	-3.309	Congestion
87	1.109	-47.46	0	1162.99	6,408	11,144	129,901	-1.228	Congestion
88	1.000	-73.17	0	8793.62	6,601	11,418	134,198	-6.396	Congestion
89	1.051	-17.25	0	355.54	6,812	11,717	138,886	-1.016	Congestion
90	1.016	0	0	0	6,995	11,987	143,019		
91	1.000	0	0	0	7,338	12,463	150,583		
92	1.000	-482.98	0	2600.43	7,463	13,426	155,543	-0.279	Congestion
93	1.000	-679.37	0	4129.32	7,557	14,147	159,257	-0.322	Congestion
94	1.030	-52.81	0	284.33	7,669	15,014	163,723	-0.254	Congestion
95	1.011	-56.77	0	305.66	7,670	15,022	163,764	-0.254	Congestion
96	1.000	0	0	0	7,822	16,191	169,786		
97	1.007	0	-105.09	226.18	7,531	16,864	168,568	-0.216	Congestion
98	1.000	0	0	0	7,201	17,627	167,187		
99	1.000	0	-921.99	1984.28	7,281	17,442	167,522	-0.232	Congestion
00	1.000	0	-2260.37	2706.52	7,053	17,438	165,554	-0.145	Congestion
01	1.000	0	0	0	6,067	16,176	154,671		
Average	1.048								

and shortfalls in the output. This finding, along with the downward trend in the scale elasticity (see Figure 5), suggests that more efforts on intense business restructuring be required to wash away diseconomies of scale. The results on our weak congestion model reveal that there is a change in the source of input congestion over time. This change occurred from the number of chain stores in late 70s and late 80s to the area of chain stores in late 90s. This model is helpful to find the sources of congestion, although it was found case sensitive in this study.

3. The analyses developed here are associated with the time series data of 27 years. Hence, the performance of a certain year is compared with all the years including later years. This might sound strange if we stick to the view that a reference is consisted only of preceding years. We can cope with this view in a way that we evaluate the year T within the data set consisting of 1, 2,... T years. However, from managerial point of view, it is useful to utilize the all past data for the evaluation of future store planning. This can be realized by adding a new plan, i.e.,

(Number, Area and Sales) to the existing data set, and then analyze the augmented data set. In this way we can evaluate the scale elasticity of the new plan, and check if it is in a status being under congestion, DRS, CRS or IRS.

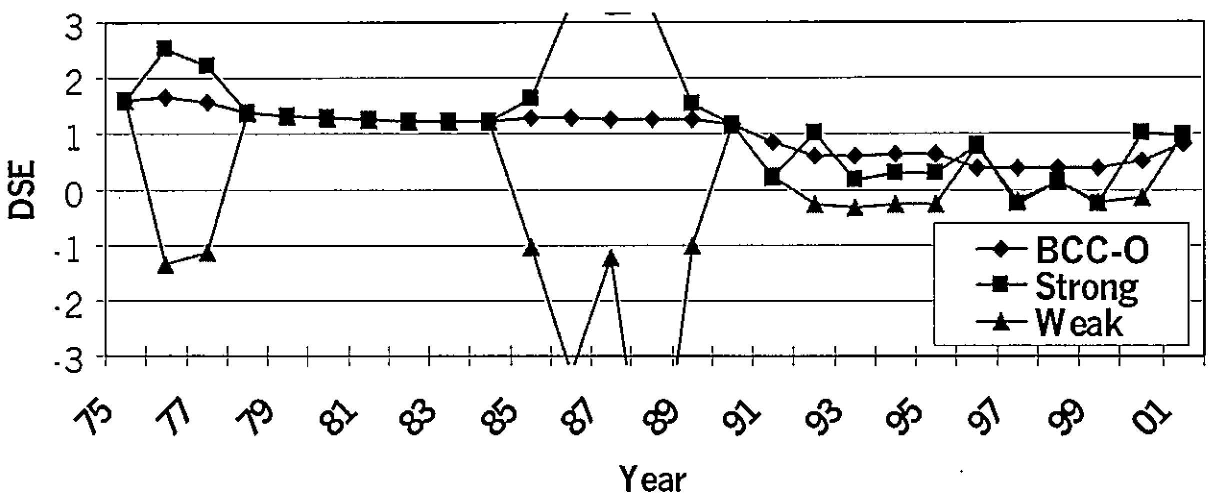

Figure 5: Scale Elasticity Comparison in Three Models

6 Concluding Remarks

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.

In this paper, we have developed a new scheme to evaluate the scale elasticity in the presence of congestion within a unified framework. We have applied this method to a data set of the Japan Chain Store Association, and found that the association had been operating under DRS for the last ten years. Even more, strong congestion was observed in the last few years, while weak congestion was found in the number of stores in 70's and 80's, and in the area of stores in 90's. These findings are helpful for managers to numerically check their future plans for restructuring their business.

To note here that 'scale economies' is meant throughout in this paper 'returns to scale', but not 'economies of scale.' However, a well distinction exists in the literature in which it is implicitly maintained that in the special case of given input factor prices, the cost structure is entirely determined from the underlying production technology where IRS implies economies of scale. However, as the input market is typically imperfect in the real world, these two concepts can no longer be the same. A description concerning the conceptual differences between these two concepts lies beyond the scope of this study. However, the interested readers can refer to our earlier studies, e.g., Sahoo *et al.* (1999) and Tone and Sahoo (2002a,b) in which both the concepts are critically analyzed and distinguished in the light of classical and neoclassical perspectives, and it is shown that they have distinctive causative factors that do not permit them to be used interchangeably.

This study points to avenues for future research. First, in any empirical study one needs to quantitatively examine the extent to which the loss (in terms of cost/profit) occurs due to congestion. Secondly, the issue of strong *vis-a-vis* weak congestion needs to be further addressed from an empirical perspective depending upon the nature of inputs used and outputs produced by the business entities in any economic environment.

References

[1] Banker, R. D., A. Charnes and W.W. Cooper. 1984. Models for the Estimation of Technical and Scale Inefficiencies in Data Envelopment Analysis. *Management Science,* 30, 1078-1092.

[2] Banker, R. D. and R.M. Thrall. 1992. Estimation of Returns to Scale Using Data Envelopment Analysis. *European Journal of Operational Re-*

search, 62: 74-84.

[3] 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.

[4] Banker, R. D., H. Chang and W. W. Cooper. 1996b. Equivalence and Implimentation of Alternative Methods for determining Returns to Scale in Data Envelopment Analysis. *European Journal of Operational Research,* 89, 473-481.

[5] Baumol, W. J., J. C. Panzar, and R. D. Willig. 1988. *Contestable Markets and The Theory of Industry Structure.* Harcourt Brace Jovanovich, New York.

[6] Brockett, P. L., W. W. Cooper, H.C. Shin and Y. Wang. 1998. Inefficiency and Congestion in Chinese Production Before and After the 1978 Economic Reform. *Socio-Economic Planning Sciences*, 32, 1-20.

[7] Charnes A., W. W. Cooper and E. Rhodes. 1978. Measuring the Efficiency of Decision Making Units. *European Journal of Operational Research* 2: 429-444.

[8] Cooper, W.W., L.M. Seiford and K. Tone. 1999. *Data Envelopment Analysis — A Comprehensive Text with Models, Applications, References and DEA-Solver Software* Norwell Mass: Kluwer Academic Publishers.

[9] Cooper, W.W., H. Deng, B.S. Gu, S.L. Li and R.M. Thrall. 2001. Using DEA to Improve the Management of Congestion in Chinese Industries (1981-1997). *Socio-Economic Planning Sciences*, 35, 227-242.

[10] Cooper, W.W., B.S. Gu and S.L. Li. 2001. Comparisons and Evaluation of Alternative Approaches to the Treatment of Congestion in DEA. *European Journal of Operational Research,* 132, 62-74.

[11] Cooper, W.W., L.M. Seiford and J. Zhu. 2000. A Unified Additive Model Approach for Evaluating Inefficiency and Congestion. *Socio-Economic Planning Sciences*, 34, 1-26.

[12] Färe, R. and L. Svensson. 1980. Congestion of Factors of Production. *Econometrica,* 48, 1745-1753.

[13] Färe, R., S. Grosskopf and C. A. K. Lovell. 1985. *The Measurement of Efficiency of Production* Boston: Kluwer Nijhoff.

[14] Färe, R., S. Grosskopf and C. A. K. Lovell. 1986. Scale Economies and Duality. *Journal of Economics,* 46, 175-182.

[15] Färe, R., S. Grosskopf and C. A. K. Lovell. 1988. Scale Elasticity and Scale Efficiency. *Journal of Institutional and Theoretical Economics,* 144, 721-729.

[16] Färe, R., S. Grosskopf and C. A. K. Lovell. 1994. *Production Frontiers* Cambridge: Cambridge University Press.

[17] Førsund, F. R. 1996. On the Calculation of Scale Elasticity in DEA Models. *Journal of Productivity Analysis,* 7, 283-302.

[18] Førsund, F. R. and L. Hjalmarsson. 2002. Are All Scales Optimal in DEA? Theory and Empirical Evidence. http://www.handels.gu.se/epc/data/html/html/PDF/gunwpe0066.pdf.

[19] Frisch, R. 1965. *Theory of Production* Dordrecht: D. Reidel Publishing Company.

[20] Fukuyama, H. 2001. Returns to Scale and Scale Elasticity in Farrell, Russell and Additive Models. *Journal of Productivity Analysis,* 16, 225-239.

[21] Grosskopf, S., D. Margaritis and V. Valdmanis. 2001. Costs of Teaching on Hospital Productivity. *Socio-Economic Planning Sciences,* 35, 189-204.

[22] Hanoch, G. 1970. Homotheticity in Joint Production. *Journal of Economic Theory,* 2, 423-426.

[23] Löthgren, M. and Tambour, M. 1996. Alternative Approaches to Estimate Returns to Scale in DEA-Models. *Stockholm School of Economics Working Paper No. 90.*

[24] Panzar, J. C. and R.D. Willig. 1977. Economies of Scale in Multi-output Production. *Quarterly Journal of Economics,* XLI, 481-493.

[25] 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.

[26] Starrett, D. A. 1977. Measuring Returns to Scale in the Aggregate, and the Scale Effect of Public Goods. *Econometrica,* 45, 1439-1455.

[27] Sueyoshi, T. 1997. Measuring Efficiencies and Returns to Scale of Nippon Telegraph & Telephone in Production and Cost Analyses. *Management Science,* 43, 779-796.

[28] Sueyoshi, T. 1999. DEA Duality on Returns to Scale (RTS) in Production and Cost Analyses: An Occurrence of Multiple Solutions and Differences Between Production-based and Cost-based RTS Estimates. *Management Science,* 45, 1593-1608.

[29] Tone, K. 2001. A Slacks-based Measure of Efficiency in Data Envelopment Analysis. *European Journal of Operational Research* 130: 498-509.

[30] Tone, K. and B. K. Sahoo. 2002a. Scale, Indivisibilities and Production Function in Data Envelopment Analysis. *GRIPS Research Report Series I-2001-0003.*

[31] Tone, K. and B. K. Sahoo. 2002b. Cost Efficiency and Returns to Scale in Life Insurance Corporation of India Using Data Envelopment Analysis. *GRIPS Research Report Series I-2002-0001.*