INTERNATIONAL CENTRE FOR ECONOMIC RESEARCH

ICER

WORKING PAPER SERIES

Finn R. Førsund and Lennart Hjalmarsson

CALCULATING THE SCALE ELASTICITY IN DEA MODELS

Working Paper No. 28 / 2002

Published in the Journal of Operational Research Society, 2004; 55 (10):1023-1038.

CALCULATING THE SCALE ELASTICITY IN DEA MODELS*

by

Finn R. Førsund

Department of Economics, University of Oslo and The Frisch Centre

and

Lennart Hjalmarsson

Department of Economics, Göteborg University

June 2002

Abstract: In economics scale properties of a production function is charcterised by the value of the scale elasticity. In the field of efficiency studies this is also a valid approach for the frontier production function. It has no good meaning to talk about scale properties of inefficient observations. In the DEA literature a qualitative characterisation is most common. The contribution of the paper is to apply the concept of scale elasticity from multi output production theory in economics to the piecewise linear frontier production function, and to develop formulas for calculating values of the scale elasticity for radial projections of inefficient observations. Illustrations also on real data are provided, showing the differences between scale elasticity values for the input- and output oriented projections and the range of values for efficient observations.

Key words: Scale elasticity, DEA, production theory, Farrell efficiency measures

JEL classification: C61, D24

---

* The paper is written as part of the Norwegian Research Council program *Efficiency in the Public Sector* at the Frisch Centre, University of Oslo. It was finished while the first author was a visiting fellow at ICER, Turin. Additional support from the following sources is gratefully acknowledged: The Bank of Sweden Tercentenary Foundation, HSFR, Jan Wallander's Research Foundation and Gothenburg School of Economics Foundation. We are indebted to Anders Hjalmarsson for carrying out all programming and calculations.

Address of corresponding author:
Box 1095 University of Oslo 0317 Blindern, Oslo, Norway
Phone: 47 2285 5127, Fax: 47 2285 5035, E-mail: f.r.forsund@econ.uio.no

1. Introduction

A core concept within economics is the efficient production function, i.e. a transformation from inputs to outputs, characterised by using the minimum of inputs for given outputs and yielding the maximum of outputs for given inputs allowed by the technology. Efficiency analysis based on a measure of distance between the efficient function and actual performance, on a micro level, started within economics with Farrell (1957). He proposed a non-parametric piecewise linear envelopment of actual performance as the estimator for the efficient - best practice - production function. Based on the discussion of his paper (Discussion, 1957) estimation methods for parametric single output frontier production functions where developed the next 20 years within economics (highlights are Aigner and Chu (1968), Afriat (1972), Meeusen and Broeck (1977), and Aigner, Lovell and Schmidt, 1977) .

Within operational research and management science the original non-parametric estimation method of Farrell were picked up and developed in Charnes, Cooper and Rhodes (1978)[1]. The term Data Envelopment Analysis (DEA) was coined there. Developments and applications of the DEA model has increased rapidly in the last decade within the OR – MS fields (see Cooper, Seiford and Tone (2000) for a recent extensive bibiliography). With the emergence of a large number of user-friendly software packages, the DEA model has now become easily accessible for practitioners. It offers a seemingly simple method for estimation of efficiency, and it accommodates easily multiple-output multiple-input technologies. Moreover, it provides a lot of useful information – not only about efficiency but also, for example, about scale properties. Indeed, one of the most frequently conducted investigations concerns returns to scale and the optimal size of decision making units (DMUs in DEA terminology).

Within the production theory in economics an efficient production function is characterised by two key features; the substitution- and scale properties. For

[1] However, similar contributions to the programming model by Berkley agricultural economists in the late 60's – early 70's were overlooked, see Førsund and Sarafoglous (2002) for an account of the history of DEA from Farrell (1957) to Charnes, Cooper and Rhodes (1978).

parametric production functions these characteristics will also be parametric functions. While there are standard procedures for analysing scale properties of analytic production functions within economics, there does not seem to be a common approach to such studies of corresponding properties of the non-parametric frontier functions within the OR-MS field. Indeed, there seems to be a lack of knowledge of production theory that has created some rather unnecessary research efforts the last decade. In Banker, Charnes and Cooper (1984) there is an expressed wish of establishing ”contacts with economics” (pp. 1079-1080). However, only production theory related to a single output case is refered to, and not much of insights about production theory in economics is really utilised. Although the case of multiple outputs is not covered as well as single output in textbooks, enough results are established to make a selective use rewarding. A seminal reference is Frisch (1965). See also Laitinen (1980) for a comprehensive review of the multiple output literature.

Standard assumptions on the general production set in the non-parametric case ensures that the substitution properies are characterised by factor- and output isoquants with the same shape as for neoclassical production functions, although the isoquants will be piecewise linear. The efficient subset of the production set is piecewise linear, and corresponds to the efficient neo-classical analytical textbook production function. It is therefore the substitution- and scale properties of the efficient subset that correspond to these properties of the efficient neoclassical production function.

We will be concerned with how to characterise scale properties of the frontier function contained in the DEA model. For a text book analytical production function the values of the scale elasticity function provide the characterisation directly. Given a parametric form of the production function the scale elasticity can be calculated and the scale characterisation done numerically. Since the DEA model is non-parametric it is not so obvious how to establish a numerical characterisation. Accordingly, most of the effort has been devoted to *qualitative* characterisation, i.e. whether the operations have increasing, constant or decreasing returns to scale. According to Banker et al. (2000, p. 26):

> ”There is a literature – albeit a relatively small one – which is directed to “quantitative” estimates of RTS [returns to scale] in DEA.”

We will explore the latter approach, following up Førsund (1996), and show rigorously how the scale elasticity can be expressed as an analytical function in the case of DEA, and how it can be calculated. The exact numerical value of the scale elasticity will of course give more information than just a qualitative characterisation of belonging to three groups. If the computational effort involved in numerical calculations also is less than what is needed doing the qualitative characterisation, then there is no reason to continue the qualitative route.

The literature will be reviewed in Section 2, and the derivation of the scale elasticity function in the case of neo-classical multiple output-multiple inputs production function stated in Section 3. Inefficiency and the DEA model are introduced in Section 4. The main results of how to calculated the scale elasticity are established in Section 5. Numerical examples from the literature and an actual data set are presented in Section 6. Section 7 concludes.

2. The DEA scale elasticity literature

Within the DEA tradition there are two different approaches to providing returns to scale information. One approach is to establish the qualitative nature of returns to scale, i.e. classify scale properties into the three categories increasing, constant and decreasing returns to scale. Another approach (presented first in Banker, Charnes and Cooper, 1984) is to establish the numerical value of the scale elasticity. The qualitative approach can be pursued along three different routes. One is due to Banker (1984) and is based on the constant returns to scale model (as a diagnostic device only, the technology must, of course, be variable returns to scale). The sum of weights defining the reference point on the frontier for each unit is used as a qualitative indicator. The second approach is due to Färe, Grosskopf and Lovell (1983) and (1985), and Färe and Grosskopf (1985), and is based on a comparison efficiency scores based on three different technology specifications, constant, non-increasing and variable returns to scale (see Grosskopf (1986) for an exposition of the nature of production set specifications in DEA models). The third approach due to Banker, Charnes and Cooper (1984) is based on inspecting the shadow price on the convexity

constraint when setting up a variable returns to scale specification.

The problem with the Banker (1984) approach was that there might be multiple solutions of the linear programming models invalidating his estimator. Subsequent research has developed procedures for handling multiple solutions when doing qualitative classification of returns to scale (see e.g. Zhu and Shen (1995), Seiford and Zhu (1998), Thore (1996), Banker, Bardhan, and Cooper (1996), Banker, Chang, and Cooper (1996), Golany and Yu, 1997). It has been established that, when taking care of the possibility of multiple solutions, the Banker (1984) classification approach and the Färe, Grosskopf and Lovell (1983) approach lead to identical classifications (see e.g. Banker et al. (2000), Seiford and Zhu (1999a), Sueyoshi, 1999). Tone (1996) provides a qualitative characterisation of all facets based on the classification of the peers spanning them.

Banker and Thrall (1992) showed how the classification of the scale elasticity could be done facing the possibility of multiple optimal solutions for the shadow price on the convexity constraint. This approach has been followed up in Førsund (1996) showing the connection to standard neo-classical production theory, in Tone (1996), in Golany and Yu (1997), and in the related papers of Banker, Bardhan, and Cooper 1996), Banker, Chang and Cooper (1996), Banker et al. (2000), and Seiford and Zhu (1999b). Sueyoshi (1997) and (1999) connects the numerical calculation of returns to scale based on the production function to the calculation based on the cost function utilising duality.

Sensitivity of the returns to scale classification was addressed in Golany and Yu (1997) and followed up in Seiford and Zhu (1999a).

In two recent studies of establishing numerical value of scale elasticity Fukuyama (2000) and (2001) set forth to expose the mathematical structure of the scale elasticity for the Farrell model and also additive models. Some new insights are achieved, but unfortunately the issue becomes quite confused because of insisting on computing scale elasticity for inefficient observations. Therefore, further elaboration still seems to be useful in this field.

3. Neoclassical production theory

The general starting point is a standard neoclassical production function $F(y,x) = 0$ for multiple outputs and multiple inputs. The output-vector is $y = (y_1,..,y_M) \in R_+^M$ and the input-vector $x = (x_1,..,x_N) \in R_+^N$:

$$F(y,x) = 0\,,\ \frac{\partial F(y,x)}{\partial y_m} > 0\,,\ m = 1,..,M\,,\ \frac{\partial F(y,x)}{\partial x_n} < 0\,,\ n = 1,..,N \qquad (1)$$

The transformation function $F(y,x) = 0$ represents the efficient output-input combinations, and it is assumed to be continuously differentiable and strictly increasing in outputs and decreasing in inputs.

The scale elasticity

The returns to scale, or scale elasticity, or the passus coefficient in the terminology of Frisch (1965), is a measurement of the increase in output relative to a proportional increase in all inputs, evaluated as marginal changes at a point in output – input space. In a multi-output setting the increase in a single output is most naturally substituted with a *proportional* increase in all outputs (see Hanoch (1970), Starrett (1977) and Panzar and Willig, 1977). Expand inputs proportionally with factor m, and pick the proportional expansion, $b = b(m, y, x)$ (with $b(1, y, x) = 1$), of outputs allowed by the transformation function:

$$F(b(m,y,x)y, mx) = 0 \qquad (2)$$

The scale elasticity, e, as a function of outputs and inputs is defined for a differentiable function as the marginal change in the output expansion factor by a marginal change in the input expansion factor over the average ratio:

$$e(y,x) = \frac{\partial b(m,y,x)}{\partial m}\frac{m}{b} \qquad (3)$$

The rule for calculating the scale elasticity is obtained by differentiating (2) with respect to the input scaling factor:

$$\frac{\partial F(by,mx)}{\partial m} = \sum_{m=1}^{M}\frac{\partial F(by,mx)}{\partial (by_m)}y_m\frac{\partial b}{\partial m} + \sum_{n=1}^{N}\frac{\partial F(by,mx)}{\partial (mx_n)}x_n = 0$$

Evaluating the derivatives, without loss of generality, at $b = m = 1$ and solving for the

scale elasticity yields[2]:

$$\frac{\partial b(y,x)}{\partial m} \equiv e(y,x) = -\frac{\sum_{n=1}^{N} \frac{\partial F(y,x)}{\partial x_n} x_n}{\sum_{m=1}^{M} \frac{\partial F(y,x)}{\partial y_m} y_m} \tag{4}$$

Equation (4) is the generalisation of Frisch`s *Passus Equation*, or sometimes called the *Generalised Euler Equation* expressing a local homogeneity property, to multiple outputs.

The economic significance of the scale elasticity

The interest in scale characterisations in production theory stems from the connection between the scale elasticity and conditions for competitive behaviour. Consider the multi- output and input production function (1), and let p_m (m =1,..,M) be the prices on outputs and q_n (n =1,..,N) be the prices on inputs, and assume that the firm operates in competitive markets both for outputs and inputs. The Lagrangian for the profit maximising problem is:

$$L = \sum_{m=1}^{M} p_m y_m - \sum_{n=1}^{N} q_n x_n - \lambda F(y,x) \tag{5}$$

We then have the following first order conditions:

$$\begin{aligned} \frac{\partial L}{\partial y_m} &= p_m - \lambda \frac{\partial F}{\partial y_m} = 0 \\ \frac{\partial L}{\partial x_n} &= -q_n - \lambda \frac{\partial F}{\partial x_n} = 0 \end{aligned} \tag{6}$$

Inserting the first-order conditions in the expression for the maximised profit yields:

$$\begin{aligned} \pi = \sum_{m=1}^{M} p_m y_m - \sum_{n=1}^{N} q_n x_n &= \sum_{m=1}^{M} \lambda \frac{\partial F}{\partial y_m} y_m - \sum_{n=1}^{N} -\lambda \frac{\partial F}{\partial x_n} x_n = \\ \sum_{m=1}^{M} \lambda \frac{\partial F}{\partial y_m} y_m \left(1 - \frac{\lambda \sum_{n=1}^{N} -\frac{\partial F}{\partial x_n} x_n}{\lambda \sum_{m=1}^{M} \frac{\partial F}{\partial y_m} y_m} \right) &= \sum_{m=1}^{M} p_m y_m \left[1 - e(y,x) \right] \end{aligned} \tag{7}$$

To derive the last expression in (7) the definition of the scale elasticity, $e(y,x)$, is

[2] To our knowledge this formula was first stated, somewhat surprisingly, as late as in Hanoch (1970).

used, and the first order conditions for outputs in (6). Maximised profits can be expressed as total revenue multiplied by a function of the scale elasticity such that profit is only non-negative for the scale elasticity smaller or equal to one. Scale elasticity values greater than one are not compatible with profit maximisation and competitive markets. This is a generalisation of a rule expressed in Frisch (1965) in the case of a single output.

The value 1 of the scale elasticity is of especial importance. We see from (7) that the profit is zero for $e = 1$. This is the condition for long-run competitive equilibrium; pure profit is exhausted. The economic benefit of such a state may be appreciated by noting a special feature of partial productivities when the scale elasticity is 1. We first have to introduce a regularity condition to ensure the text-book S-shape of the production function and U-shaped average cost curves. We will use the (generalised) *Regular Ultra Passum* law of Frisch (1965) (see Førsund and Hjalmarsson (2002) for an exposition). Due to monotonicity of the production function all movements in input-output space satisfying the condition that none of the outputs and inputs are decreasing and at least one output and one input are increasing, must then pass through unique points where the scale elasticity is 1. If we as a special case consider *proportional* variation in outputs with the factor b and variations in the inputs with factor m for points satisfying the production function (1), then an important result is that productivity, defined as the ratio b/m, along a ray is *maximal* at the point where the scale elasticity equals 1:

$$Max_m \frac{b(m,y,x)}{m} \rightarrow \frac{\partial(b/m)}{\partial m} = \frac{\frac{\partial b}{\partial m}m - b}{m^2} = 0 \Rightarrow \frac{\partial b}{\partial m}\frac{m}{b} \equiv e = 1 \qquad (8)$$

using the definition (3) of the scale elasticity. The long-run competitive equilibrium also ensures that the resources are utilised most productively. The scale of an operation for $e = 1$ is in Frisch (1965) termed *Technically Optimal Scale* (shortened to *TOPS* in Førsund and Hjalmarsson, 2002)[3].

[3] In the OR-MS literature this classical concept in production theory is overlooked completely, and the identical term MPSS (most productive scale size) in Banker (1984) is cited as an original concept.

4. Introducing inefficiency

So far efficient operations have been assumed. In order to deal with inefficient operations we need a production technology where both feasible efficient and inefficient points can be identified. A production possibility set *S* is in general defined by:

$$S = \{(y, x) : x \, can \, produce \, y\} \tag{9}$$

Basic regularity conditions assumed for *S* is that the output- and input vectors are drawn from bounded sets, *S* includes its limit points, positive production cannot occur without positive inputs, and free disposals of inputs and outputs (increase in inputs must lead to increased or constant outputs, and a smaller output vector than a feasible vector is also feasible, employing the same inputs, see e.g. Färe and Primont, 1995). We need to distinguish between efficient and inefficient points as subsets of the production set *S*. The connection between the neoclassical production function (1) and the production set formulation (9) is as follows (see Hanoch (1970), and McFadden (1978), which states conditions for a unique connection), with the standard properties on *S* as stated above:

$$S = \{(y, x) : x \, can \, \, produce \, y\} \equiv \{(y, x) : F(y, x) \leq 0\} \tag{10}$$

The subset of efficient point is then defined by $F(y,x) = 0$.

Efficiency measures

Inefficiencies are measured by efficiency scores as defined by Farrell (1957) and extended to variable returns to scale in Førsund and Hjalmarsson (1974) and (1979b), and Färe and Lovell (1978). An inefficient observation can be related to the frontier technology through potential input saving or potential output expansion. In the first case by a proportional shrinking of inputs, the input saving measure, E_1; in the second case by expanding observed outputs proportionally to the frontier using observed inputs on frontier technology, the output-increasing measure, E_2.[4] The input saving

[4] Farrell (1957) used the lower case notation e_1 and e_2.

efficiency measure is:

$$E_1(y,x) = Min_q \{q : (y,qx) \in S\} \equiv Min_q \{q : F(y,qx) = 0\} \quad (11)$$

The output increasing measure is:

$$E_2(y,x) = Min_f \left\{ f : (\frac{y}{f}, x) \in S \right\} \equiv Min_f \left\{ f : F(\frac{y}{f}, x) = 0 \right\} \quad (12)$$

The E_1 measure is identical to the inverse of the Shephard input distance function, and the E_2 measure is identical to the Shephard output distance function. Both measures are conditional upon the production possibility set, or the efficient production function. We assume enough smoothness so that the minimum is defined and unique. Notice that the efficiency measures are general as to production technology and not specific to the DEA model[5].

The DEA model

Following Farrell (1957) the production possibility set, *S*, is empirically defined by enveloping the observations as tightly as possible by a piecewise linear outer boundary (see e.g. Banker,Charnes and Cooper (1984) for the properties of this empirically based set *S*):

$$S = \{(y,x) : \sum_{j=1}^{J} l_j y_{mj} \geq y_m \ \ (m = 1,..,M), \sum_{j=1}^{J} l_j x_{nj} \leq x_n \ (n = 1,..,N), \quad \sum_{j=1}^{J} l_j = 1, l_j \geq 0 \ (j = 1,..,J)\} \quad (13)$$

There are *J* observations and the non-negative weights, l_j, determine the refererence points on the frontier. Restricting the sum of weights to be 1 implies variable returns to scale (VRS).

It has become a common practice in the field of non-parametric efficiency analysis to name the linear programme for the calculation of Farrell (1957) technical efficiency scores for the *DEA model*. The efficiency scores (11) and (12) for the VRS input- and output oriented DEA models, E_{1i} and E_{2i} respectively for unit *i*, are found by solving the following two linear programmes:

[5] In fact the Farrell efficiency measures were extended to variable returns to scale in Førsund and Hjalmarsson (1974) and (1979b), and in Färe and Lovell (1978).

$$
\begin{aligned}
& E_{1i} = Min\ q_i \\
& s.t. \\
& \sum_{j=1}^{J} l_j y_{mj} - y_{mi} \geq 0\ ,\ m = 1,..,\ M \\
& q_i x_{ni} - \sum_{j=1}^{J} l_j x_{nj} \geq 0\ ,\ n = 1,..,\ N \\
& \sum_{j=1}^{J} l_j = 1 \\
& l_j \geq 0\ ,\ j = 1,..,\ J
\end{aligned}
\tag{14}
$$

$$
\begin{aligned}
& 1/E_{2i} = Max f_i \\
& subject\ to \\
& f_i y_{mi} - \sum_{j=1}^{J} l_j y_{mj} \leq 0\ , m = 1,..,M \\
& \sum_{j=1}^{J} l_j x_{nj} - x_{ni} \leq 0\ , n = 1,..,N \\
& \sum_{j=1}^{J} l_j = 1 \\
& l_j \geq 0\ ,\ j = 1,..,J
\end{aligned}
\tag{15}
$$

For notational ease the unit index *i* is suppressed on the l- weights and the same symbols are used for the l– weights in (13), (14) and (15). The constraints in (14) and (15) represent the definition of the piecewise linear technology relevant for unit *i*. This unit may be inefficient in e.g. its use of inputs. The input vector in (14) for unit *i* is adjusted by the efficiency score, q_i, and then compared with the *reference point*, $\sum_{j=1}^{J} l_j x_{nj}$, on the frontier. The output vector in (15) for unit *i* is marked up with the factor f_i, and then compared with the reference point $\sum_{j=1}^{J} l_j y_{mj}$.

We will also need the Lagrangians, L_1 and L_2, associated with (14) and (15), set up in such a way that the shadow prices, u_{mi} and v_{ni}, on the inequality constraints for outputs and inputs respectively are non-negative:

$$L_1 = q_i - \sum_{m=1}^{M} u_{mi}(\sum_{j=1}^{J} l_j y_{mj} - y_{mi}) - \sum_{n=1}^{N} v_{ni}(q_i x_{ni} - \sum_{j=1}^{J} l_j x_{nj}) - u_i^{in}(\sum_{j=1}^{J} l_j - 1) \qquad (16)$$

$$L_2 = f_i - \sum_{m=1}^{M} u_{mi}(\sum_{j=1}^{J} -l_j y_{mj} + f_i y_{mi}) - \sum_{n=1}^{N} v_{ni}(-x_{ni} + \sum_{j=1}^{J} l_j x_{nj}) - u_i^{out}(\sum_{j=1}^{J} l_j - 1) \qquad (17)$$

Again, the same symbols are used for the shadow prices in (16) and (17). The non-restricted variables, u_i^{in} *and* u_i^{out}, are the shadow prices on the equality constraint on the sum of the l's.

5. The scale elasticity in the DEA model

Within neo-classical production theory the scale elasticity is simply determined for estimated functional forms by applying (4). When the production function is on a parametric form this is straightforward[6]. A problem in the DEA model is that since the production function is non-parametric, we cannot derive the scale elasticity as a parametric function based on the production function using (4). Another problem in the DEA model is the existence of inefficient points. It should be born in mind that returns to scale is a local property and applies only to *efficient* points, i.e. points satisfying *F(y,x)* = 0. To associate an *inefficient* point with a scale elasticity value is at best ambiguous, because the existence of inefficieny means that the local increase in outputs when inputs are increased cannot be separted from the increase due to a reduction in inefficiency[7]. Therefore, a very basic observation for the discussion of scale properties using the DEA model is that inefficient observations must first be represented by efficient points. Thus the discussion of scale properties for inefficient units must be conditional on a meaningful and interesting representation on the

[6] In the first empirical estimation of a multi output function, according to Laitinen (1980), Klein (1952) used the function $y_1 y_2^{-d} - A\Pi_{s=1}^{4} x_s^{a_s} = 0$ (in our symbols). Applying Eq. (4) then yields the scale elasticity function $e(y,x) = \sum_{s=1}^{4} a_s /(1-d)$. Its properties can then be checked analytically. The scale elasticity is constant, i.e. the production function is homogeneous. In Førsund and Hjalmarsson (1979a) a parametric homothetic *frontier* function is estimated. The scale elasticity is then constant along an isoquant.

[7] Banker (1984) and Banker, Charnes and Cooper (1984) are clear on this point. However, notice that a set is usually defined as having constant returns to scale if all finite points on rays belong to the set, i.e. the set is a cone. The definition of economies of scale in Panzar and Willig (1977) as a property of

frontier. In the DEA literature the input- or output-oriented reference points defined after (14) and (15), or the radial projections, are often called ”target values” (see e.g. Thanassoulis, 2001). However, target values must be conditional upon economic conditions, like the objective of the unit, the markets it faces, technical possibilities of realising frontier techniques, time involved, etc. It makes little sense to formulate targets without paying attention to such underlying features. It should also be remembered that the DEA model does not explain *why* a unit is inefficient, but only provides measures of distances from the best practice frontier. However, in order to obtain a scale property reference for inefficient units it has been common to use the input- and output-oriented projections. We will follow this practice, but base our approach on *radial* projections. The reason why we deviate from most of the literature and do not apply the reference points as points of projections (i.e. include slacks on the relevant constraints in addition to radial change), will be explained later.

The point of departure is the approach to characterise scale properties first introduced in Banker, Charnes and Cooper (1984). We then need the dual programmes to the problems (14) and (15).

The dual programmes

Using the shadow price variables introduced in (16) and (17) the dual problems for the primals (14) and (15) are:

$$Max \left\{ \sum_{m=1}^{M} u_{mj} y_{mj} + u_i^{in} \right\}$$

subject to

$$\sum_{n=1}^{N} v_{ni} x_{ni} = 1 \qquad (18)$$

$$\sum_{m=1}^{M} u_{mj} y_{mj} - \sum_{n=1}^{N} v_{nj} x_{nj} + u_i^{in} \leq 0 \ , \ j = 1,..,J$$

the production *set* in general is rather awkward.

$$Min \left\{ \sum_{n=1}^{N} v_{ni} x_{ni} + u_i^{out} \right\}$$

subject to

$$\sum_{m=1}^{M} u_{mi} y_{mi} = 1 \tag{19}$$

$$- \sum_{m=1}^{M} u_{mj} y_{mj} + \sum_{n=1}^{N} v_{nj} x_{nj} + u_i^{out} \geq 0 \ , \ j = 1,.., J$$

As pioneered in Banker, Charnes and Cooper (1984) the shadow prices u_i^{in} and u_i^{out} on the convexity constraint can be used to characterise the scale properties. But as noted before it is only meaningful to characterise scale for efficient points. Let us first study observations classified as efficient, i.e. units obtaining efficiency scores of 1 and having positive (or zero) shadow prices on output- and input constraints (i.e. zero slacks in general). Efficient units are by definition vertice points or located on ridges delineating facets. This is also the case for reference points when slacks occur on the constraints in (14) and (15). Consequently we do not have differentiability at such points in general, and the scale elasticity calculation (4) cannot be applied. As observed by Banker, Charnes and Cooper (1984) the solutions for u_i^{in} and u_i^{out} are not unique for efficient units. For these we know that the objective functions in (18) and (19) must be equal to 1. For $j = i$ in the last constraint of both duals we then see that the constraints must hold with equality *independently* of the values of u_i^{in} and u_i^{out}. We may then have multiple solutions for these shadow prices. Intuitively, since an efficient unit may belong to several facets we may have several returns to scale characteristics associated with points infinitesimally close to an efficient observation. What can be done is to investigate the range of the feasible solutions for the points. If the range covers zero values we know that constant returns to scale (CRS) is a possibility, since zero shadow prices on the equality constraint on the sum of l_j's being 1 means that this constraint is not binding in the primal problems (14) and (15), i. e. we are back to the CRS case. We will return to the determination of the range for u_i^{in} and u_i^{out} below.

Scale elasticity in DEA

The values of the shadow prices u_i^{in} and u_i^{out} are only unique for *inefficient* units in general. The question is how these shadow prices may be used for calculating scale elasticity values. We will proceed by a radial projection of the inefficient points to the frontier, i.e. using the efficiency scores obtained from the primal programmes (14) and (15). The reason why we are not following the common procedure (see e.g. Banker et al., 2000) of using the reference points as projections (i.e. in addition to radial change also considering the slacks on output- and input constraints in the problems) is that as observed above we do not have differentiability at such points in general since they are vertices or located at ridges delineating the facets. Radial projections will in general be interior points on the facets[8].

PROPOSITION 1

Consider an efficient hypothetical observation, i($i \in I$, set of inefficient observations), characterised by the output- and input vectors (y_i, x_i^*), with $x_i^* = E_{1i}\, x_i$, where the input-oriented efficiency score, E_{1i}, is defined by (11) and calculated by solving (14), and (y_i, x_i) is an inefficient unit with $E_{1i} < 1$. Assume that the projected point is an interior point on a facet. We can then state the following results:

a) The scale elasticity, defined by (4), for the efficient hypothetical observation (y_i, x_i^*) can be calculated as:

$$e(y_i, E_{1i}x_i) = \frac{E_{1i}}{E_{1i} - u_i^{in}}, \; i \in I \tag{20}$$

where u_i^{in} is the shadow price on the equality constraint $\sum_{j=1}^{J} l_j = 1$ calculated by solving the dual programme (18).

b) The radial projected observation (y_i, x_i^*) to the DEA frontier exhibits increasing returns to scale if $u_i^{in} > 0$, constant returns to scale if $u_i^{in} = 0$, and decreasing returns to scale if $u_i^{in} < 0$.[9]

[8] Radial projections and reference points coincide when all slacks are zero.

[9] Note that the sign convention is dependent on how the dual programme is set up.

PROOF:

Part a): The scale elasticity is defined in (4) in terms of partial derivatives of the frontier function *F(y,x)*. Since the efficiency measure functions (11), (12) give complete representations of the technology set, we have that $E_1(y, x) - 1 = 0$, $E_2(y, x) - 1 = 0$ and $f(y,x) - 1 = 0$ must give the same representation of the frontier as $F(y,x) = 0$ for the efficient points (x,y) (see e.g. McFadden (1978), Hanoch, 1970). We may then utilise this equivalence and establish the scale elasticity function following the procedure leading to (4) for either $E_1(y, x)$, $E_2(y, x)$ or $f(y,x)$ (cf. McFadden, 1978). Applying the scale elasticity expression (4) for the projected observation, (y_i, x_i^*), and writing $E_{1i}^*(.)$ for the corresponding efficiency function we have:

$$e(y_i,q_i x_i) = -\frac{\sum_{n=1}^{N} \frac{\partial E_{1i}^*(y_i, x_{ni}^*)}{\partial x_{ni}^*} x_{ni}^*}{\sum_{m=1}^{M} \frac{\partial E_{1i}^*(y_i,q_i x_i)}{\partial y_{mi}} y_{mi}} = \frac{1}{\frac{1}{q_i}\sum_{m=1}^{M} \frac{\partial E_{1i}(y_i, x_i)}{\partial y_{mi}} y_{mi}} = \frac{1}{\frac{1}{q_i}\sum_{m=1}^{M} u_{mi} y_{mi}} =$$

$$\frac{1}{\frac{1}{q_i}(q_i - u_i^{in})} = \frac{E_{1i}}{E_{1i} - u_i^{in}} \qquad (21)$$

We first apply the property of homogeneity of degree –1 in inputs to derive the expression after the second equality sign. The numerator is obtained applying the Euler Theorem to homogeneous functions, or also termed the *Passus Equation* for a single output (i.e. E_{1i}^*) "production function" in Frisch (1965), as mentioned after (4), remembering that $E_{1i}^* = 1$[10]. Applying the homogeneity property again we get the denominator. Notice the switch from the $E_{1i}^*(.)$ function to the $E_{1i}(.)$ function. This is a crucial step. The denominator in the expression after the third equality sign is obtained by applying the Envelope Theorem to problem (14) for the inefficient observation (y_i, x_i). Using the Lagrange function (16) we get:

[10] We could also use that $-\sum_{n=1}^{N} \frac{\partial E_{1i}^*(y_i, x_{ni}^*)}{\partial x_{ni}^*} x_{ni}^* = \sum_{n=1}^{N} v_{ni}^* x_{ni}^* = 1$, where the second expression is obtained using the envelope theorem yielding $\frac{\partial E_{1i}^*}{\partial x_{ni}^*} = \frac{\partial L_1}{\partial x_{ni}^*} = -v_{ni}^*$, remembering that we are at an efficient point with $E_{1i}^* = 1$. The third expression is obtained from the corresponding dual programme (18) where we have from the first constraint that $\sum_{n=1}^{N} v_{ni}^* x_{ni}^* = 1$.

$$\frac{\partial q_i}{\partial y_{mi}} = \frac{\partial L_1}{\partial y_{mi}} = u_{mi} \tag{22}$$

From the objective function of the dual programme (18) we have $\sum_{m=1}^{M} u_{mi} y_{mi} = q_i - u_i^{in}$, yielding the denominator in the expression after the fourth equality sign. For the final expression we substitute E_{1i} for q_i.

We may establish (21) in a slightly shorter way by noting, following Caves, Christensen and Diewert (1982), that the definition of scale the elasticity may be obtained by asking the question of establishing the minimal proportional increase, m, in inputs corresponding to a proportional increase, b, in outputs; i.e. defining the scale elasticity as $e = 1/(\partial m/\partial b)$ (evaluated at $b = 1$). Noting from (2) and (11) that the scaling factor $m(by, x)$ (with $m(y, x) = 1$) corresponds to the input-saving efficiency function $E_1(by, x)$ we have immediately:

$$e(y_i, x_i^*) = 1/\sum_{m=1}^{M} \frac{\partial E_{1i}^*(y_i, x_i^*)}{\partial y_{mi}} y_{mi} \tag{23}$$

We can then substitute for x_i^* and use the homogeneity property of the $E_{1i}^*(.)$–function and apply the Envelope Theorem and use the dual (18) as above.

REMARK 1

From the objective function of the dual (18) the maximal value of u_i^{in} is 1 when a unit is efficient, since the shadow prices on the output constraints are non-negative. The corresponding value of the scale elasticity is then infinity for an efficient unit. The minimal value of u_i^{in} is minus infinity for an efficient unit. The corresponding value of the scale elasticity is then zero. This range shows the non-uniqueness of this shadow price for efficient (original) units.

The scale elasticity may also be calculated based on a radial output-oriented projection of an inefficient unit.

PROPOSITION 2.

Consider an efficient hypothetical observation, i ($i \in I$, set of inefficient observations),

characterised by the output- and input vectors (y_i^*, x_i), with $y_i^* = (1/E_{2i})y_i$, where the output-oriented efficiency score, E_{2i}, is defined by (12) and calculated by solving (15), and (y_i, x_i) is an inefficient unit with $E_{2i} < 1$. Assume that the projected point is an interior point on a facet. We can then state the following results:

a) The scale elasticity, defined by (4), for the efficient hypothetical observation (y_i^*, x_i) can be calculated as:

$$e(\frac{y_i}{E_{2i}}, x_i) = 1 - E_{2i}(y_i, x_i)u_i^{out},\ i \in I \tag{24}$$

where u_i^{out} is the shadow price on the equality constraint $\sum_{j=1}^{J} l_j = 1$ calculated by solving the dual programme (19).

b) The radial projected observation (y_i^*, x_i) to the DEA frontier exhibits increasing returns to scale if $u_i^{out} < 0$, constant returns to scale if $u_i^{out} = 0$, and decreasing returns to scale if $u_i^{out} > 0$ (cf. footnote 9).

PROOF:

Part a): Proceeding as for the proof of proposition 1 we have the scale elasticity expression for the efficient hypothetical observation (y_i^*, x_i), using the equivalence between the general frontier production function *F(y,x)* and the output-oriented mark-up function *f*(y,x)* for an efficient observation:

$$e(f_i y_i, x_i) = -\frac{\sum_{n=1}^{N} \frac{\partial f_i^*(f_i y_i, x_i)}{\partial x_{ni}} x_{ni}}{\sum_{m=1}^{M} \frac{\partial f_i^*(y_i^*, x_i)}{\partial y_{mi}^*} y_{mi}^*} = \frac{\frac{1}{f_i}\sum_{n=1}^{N} \frac{\partial f_i(y_i, x_i)}{\partial x_{ni}} x_{ni}}{1} =$$

$$\frac{\frac{1}{f_i}\sum_{n=1}^{N} v_{ni} x_{ni}}{1} = \frac{\frac{1}{f_i}(f_i - u_i^{out})}{1} = 1 - E_{2i}u_i^{out},\ i \in I \tag{25}$$

In the expression after the second equality sign we have in the numerator utilised that the f_i*(.) function is homogeneous of degree –1 in outputs, and in the denominator we have again utilised the Euler Theorem for homogeneous functions, or the Passus

Equation[11]. To derive the numerator in the expression after the third equality sign the Envelope Theorem is applied to (19):

$$\frac{\partial f_i}{\partial x_{ni}} = \frac{\partial L_2}{\partial x_{ni}} = v_{ni} \tag{26}$$

In the expression after the fourth equality sign the dual programme (19) is used for observation *i* deriving the expression in the numerator, and for the final expression E_{2i} is substituted for $1/f_i$.

Part b) is established by observing the size of the scale elasticity in (25) following from the three possible states of the shadow price u_i^{out}.

As for Proposition 1 the expression (25) may be derived somewhat more directly by observing, comparing (12) and (2), that the mark-up factor f_i plays the same role as the proportional expansion factor $b = b(y, mx)$ when seeking the maximal proportional change, b, in outputs for a proportional change, m, in inputs in the definition (3) of the scale elasticity:

$$e(f_i y_i, mx_i) = e(f_i y_i, mx_i) = \frac{\partial f_i^*}{\partial m} = \sum_{n=1}^{N} \frac{\partial f_i^*(f_i y_i, x_i)}{\partial x_{ni}} x_{ni} \tag{27}$$

where the derivatives are evaluated at m =1. From this stage on we continue as in the derivation of (25).

REMARK 2

Since the shadow prices on the input constraints are non-negative we have from (19) that the maximal value of u_i^{out} is 1 when a unit is efficient. The corresponding value of the scale elasticity is then zero for an efficient unit. The minimal value of u_i^{out} is minus infinity. The corresponding value of the scale elasticity is then plus infinity. Again we see the potential for multiple solutions for the shadow price.

---

[11] We could also use that $-\sum_{m=1}^{M} \frac{\partial f_i^*(y_i^*, x_i)}{\partial y_{mi}^*} y_{mi}^* = \sum_{m=1}^{M} u_{mi}^* y_{mi}^* = 1$, where the second expression is obtained using the envelope theorem yielding $\frac{\partial f_i^*}{\partial y_{mi}^*} = \frac{\partial L_2}{\partial y_{mi}^*} = -u_{mi}^*$, remembering that we are at an efficient point with $f_i^* = 1$. The third expression is obtained from the corresponding dual programme (19) where we have from the first constraint that $\sum_{m=1}^{M} u_{mi}^* y_{mi}^* = 1$.

Propositions 1 and 2 provide simple expressions for a numerical calculation of efficient hypothetical points being radial projections of inefficient units when solving the input – and output oriented DEA model (14) and (15). For the calculation of the scale elasticity, only the shadow prices on the convexity constraint is needed together with the efficiency scores. These are remarkably simple formulas compared with Equation (4), which requires one to evaluate sums of partial derivatives. Furthermore, we do not have to solve any new program, assuming that it is standard to get the shadow prices when solving the primal linear programming problems (14) and (15).

The two scale elasticity values for the same inefficient observation resulting from applying the rules in Propositions 1 and 2, may be seen as representing the most acceptable range for efficient hypothetical projections, in the sense that no changes in outputs are required in the input-oriented case and only reductions required in inputs keeping the mix fixed, while no changes are required in inputs and only increases in outputs, keeping the mix fixed, for the output-oriented case. It may be asked if an average characterisation can be established for this range. Such a characterisation was provided in Førsund and Hjalmarsson (1979b) for single output and generalised to multiple outputs in Førsund (1996), and is included here for completeness.

PROPOSITION 3

Consider the radial projections $(y_i, E_{1i}x_i)$ and $((1/E_{2i})y_i, x_i)$ for an inefficient unit i ($i \hat{I} I$), where E_{1i} and E_{2i} are calculated by solving (14) and (15). We then have the following relationship between the efficiency measures and the average scale elasticity, $\bar{e}$, over the path from the input-oriented to the output-oriented projection (suppressing the unit index i):

$$E_2(\mathrm{y},\mathrm{x}) = E_1(\mathrm{y},\mathrm{x})^{\bar{e}} \tag{28}$$

The average scale elasticity, $\bar{e}$, is defined by:

$$\bar{e} = \frac{\int_{m=1}^{m} e(y, mx)\, d\ln m}{\int_{m=1}^{m} d\ln m} \tag{29}$$

Sketching the derivation, starting at the input-oriented projected point, (y, E_1x), on the frontier, we change the inputs with the factor m and the output with factor $b(y,mx)$ ($b(y,x)$ =1) until the output-oriented point on the frontier, $((1/E_2)y, x) = (by, mE_1x)$, is

reached, assuming that such a path is feasible. The value of m at the output-oriented projection is m. Associating $1/E_2$ with b and $1/E_1$ with m ($x = mE_1x \Rightarrow m = 1/E_1$) we get (29) by using the generalised *Second form of the Beam Variation Equation* in Frisch (1965), see Førsund (1996) for the complete proof.

The scale elasticity for efficient units

As observed in Banker, Charnes and Cooper (1984) and Banker and Thrall (1992) the solution for the shadow price on the convexity constraint in problems (14) or (15) may not be unique, making the reference to supporting hyperplanes of the frontier. By the nature of a piecewise linear frontier, the efficient observations must in general represent vertices. Marginal productivities and scale elasticities are not defined for such points. The standard procedure is to evaluate left- and right-hand derivatives[12]. In the DEA model an efficient unit may be a corner point for several facets. It seems most reasonable to calculate the maximal range of the shadow price for the efficient unit in question. We have already identified extreme values in the Remarks 1 and 2.

Banker and Thrall (1992) offer a straightforward way of calculating upper and lower bounds for the scale elasticity for an efficient observation i ($i \in P$, where P is the set of efficient observations, i.e. $E_{1i}=1$ and all slacks zero). In the case of an input - adjusted frontier point, finding the maximal value of the key parameter u_i^{in} in (16) is done by substituting the objective function in the dual problem (16) with the shadow price itself and add to the two restrictions in (16) the constraint implied by the observation being efficient :

$$\begin{aligned} &Max\ u_i^{in} \\ &subject\ to \\ &\sum_{m=1}^{M} u_{mi} y_{mi} + u_i^{in} = 1 \qquad (30) \\ &\sum_{n=1}^{N} v_{ni} x_{ni} = 1 \\ &\sum_{m=1}^{M} u_{mj} y_{mj} - \sum_{n=1}^{N} v_{nj} x_{nj} + u_i^{in} \le 0\ ,\ j = 1,..,J \end{aligned}$$

[12] See Førsund and Hjalmarsson (1987) for an application calculating scale elasticity in the case of one output and two inputs in a piecewise linear production model.

Writing the solution as $u_i^{in\text{-}max}$ using (21) yields:

$$e^{\max}(y_i, x) = \frac{1}{1 - u_i^{in-\max}}, i \in P \tag{31}$$

To find the minimal bound value of the shadow price the sign in the objective function (30) is just changed from positive to negative value, i.e. $\{u_i^{in}\}$ is replaced with $\{-u_i^{in}\}$. Writing the solution as $u_i^{in\text{-}min}$, we have the minimal scale elasticity for an efficient unit:

$$e^{\min}(y_i, x_i) = \frac{1}{1 - u_i^{in-\min}}, i \in P \tag{32}$$

Since there is no difference by definition in orientation for an efficient unit the calculation of the bounds using (30) for finding maximal value and change of sign for the objective function for finding the minimal value, also applies for the output-oriented model. From (20) and (24) we have for *efficient* units:

$$e(y_i, x_i) = \frac{1}{1 - u_i^{in}} = 1 - u_i^{out} \Rightarrow u_i^{out} = -\frac{u_i^{in}}{1 - u_i^{in}}, i \in P \tag{33}$$

These relationships also hold for the extreme values. The maximal value of u_i^{in} is 1 corresponding to a minimal value $-\infty$ for u_i^{out}, and yielding the maximal value of $+¥$ for the scale elasticity. The minimal value of u_i^{in} is $-¥$, corresponding to the maximal value of 1 for u_i^{out}, yielding the minimal value of zero for the scale elasticity.

If the minus infinity value creates problems as to running a specific LP-software, we can adapt the proposal in Banker, Bardhan, and Cooper (1996) to eliminate the problem of infinite solution. If a negative value of u_i^{in} is obtained in the solution of (14), then by running the maximisation problem (30) with the objective function $\{u_i^{in}\}$ with an additional constraint of $u_i^{in} \le 0$, the maximal non-positive value is found[13]. A binding constraint is then implying that constant returns to scale is present at least at one of the facets that the unit is a member of.

[13] Note that we do not have to run the same procedure if a positive value of u_i^{in} is obtained in the first run, since we know, as discussed in Remark 1, that the maximal value is 1. This seems to be overlooked in Banker, Bardhan, and Cooper (1996). If the output-oriented problem (15) is run, then for a positive value of u_i^{out} the modified program (30) is run with $\{-u_i^{out}\}$ as objective function and $u_i^{out} \ge 0$. Again, the modified program does not have to be run for negative values.

Remarks on the literature

Key contributions in the OR-MS DEA literature are Banker, Charnes and Cooper (1984) and Banker and Thrall (1992). But notice that neither paper actually claimed explicitly to calculate the scale elasticity *numerically*. Banker, Charnes and Cooper (1984) demonstrate an equivalence between using the sign of u_i^{in} and whether a parameter (termed $\hat{S}$) is greater or smaller than one. This parameter was informally linked to the elasticity of scale in the single output case. This failure to address explicitly the calculation of the numerical values of scale elasticities is also the case in Banker and Thrall (1992), although some analogy is made with a linear single output production function. In Cooper, Thompson and Thrall (1996), Sueyoshi (1997) and Fukuyama (2000) formulas for calculating the value of the scale elasticity are *defined*, not derived[14]. In Banker, Bardhan and Cooper (1996) it is claimed that Banker and Thrall (1992) "provide a measure of scale elasticity". However, the precise rule (4) from production theory to calculate the numerical value of the scale elasticity is never considered formally. What is derived rather ad hoc in both papers as characterising the *nature* of returns to scale is the number $1/(1- u_i^{in})$, which is the rule (20) for $E_{1i} =1$.

In Banker, Bardhan and Cooper (1996) the change in the program (30) described above to avoid the solution of minus infinity (it is unnecessary to worry about plus infinity since this is not feasible when we stick to the input-orientation, which is all that is needed) is introduced as a solution to "avoid the need for exploring *all* alternate optima, .."(p. 584)[15]. However, it is hardly of interest to do this, what is needed is to find the maximal and minimal bounds, as done by the Banker and Thrall (1992) proposal.

Golany and Yu (1997) point to the problem of classifying constant returns to scale when u_i^{in} alternates in sign. However, this is not a problem to handle when calculating the scale elasticity bounds as described above including the infinity and zero values.

[14] In Sueyoshi (1997) and (1999) there seems to be a mistake in the formula, see Eq. (36) in the latter, probably due to not paying sufficient attention to keeping the efficient projection and the inefficient observation apart.

[15] To underline the usefulness of their approach Banker, Bardhan and Cooper (1996) cite Ganley and Cubbin (1992) on the difficulties of using the bounds approach in Banker and Thrall (1992). Actually the attempt at application in Ganley and Cubbin is the first one we know about in the literature. However, they have completely misunderstood the purpose of finding bounds, confusing it with finding optimal scale units. Unfortunately the book is full of misunderstandings of DEA and a warning to potential readers is hereby given.

Different signs for an efficient unit means that on the facet with the negative sign decreasing returns to scale prevails, while for the facet with positive sign increasing returns prevail. The efficient unit is then most reasonably defined to be of optimal scale, as proposed by Banker and Thrall (1992).

As stated before it has no meaning to say that an *inefficient* observation has a scale property. Nevertheless, it may seem that such language is often used in the DEA literature, may be just without making the explicit distinction between efficient projected points and the inefficient observations. However, Fukuyama (2000) and (2001) is very insistent on having derived scale elasticity expressions for inefficient observations. His way of establishing scale elasticity expressions has features in common with our derivation. The fundamental oversight right from the beginning is that there is only one frontier technology in the DEA model, and the scale elasticity is a local property of a point on this frontier; end of story. Fukuyama (2000) is in trouble already when entering the scale elasticity function in his Equation (7), because it is written as a function of an *inefficient* observation. *As if-* reasoning on parallel movements of geometric features to the interior of the technology set, made possible by homogeneity properties of the efficiency functions, must not be confused with the proper definition of the scale elasticity.

In the DEA literature there is sometimes made a distinction between global and local scale characterisations (see e.g. Banker et al., 2000). We know from the general definition in production theory in economics that the scale elasticity is a local property at a point on the production function. What seems to lie behind the expression global characterisation is that in the DEA model it is possible to give qualitative characterisations of entire facets. However, this is not in conflict with the standard approach in economics of evaluating scale properties locally. It is just that the piecewise linear structure of the non-parametric DEA frontier function makes it possible to characterise all interior local points on a facet qualitatively. From the general expression (4) for the calculation of the scale elasticity we have that all partial derivatives at a point are involved. The consequence of the piecewise linear structure is that these partial derivatives at a point can be reduced to an expression involving just the efficiency measure and the shadow price on the convexity constraint in either (14) or (15) needed for specifying variable returns to scale (VRS). As long as we are

on the same facet this shadow price is constant. It then follows that any inefficient point we may construct keeping us on the same facet when doing radial projections to the frontier will have the same qualitative characterisation as to increasing, constant or decreasing scale elasticity according the expressions stated in Propositions 1 and 2. These propositions may then also be used for "global" characterisations. Theorem 3 in Tone (1996) is another way of expressing this feature of a facet.

6. Pedagogical illustrations

Data illustrations used in the literature

Before turning to real data it may be helpful to apply the propositions (20), (24) and (28) to very simple constructed data sets used in the literature for the purpose of illustration of various properties of the DEA model.

In Banker and Thrall (1992) some efficient units in the single output – single input case was used for illustrations. Running our programme[16] we obtain the results set out in Table 1. All classification are in accordance with Banker and Thrall (1992), except

Table 1. Data from Banker and Thrall (1992)

Unit	Output	Input	ε^{max}	ε^{min}
1	1	1	∞	5
2	3.5	1.5	2.14	2.14
3	6	2	1.67	0.67
4	7	2.5	0.71	0.71
5	8	3	0.75	0.38
6	9	4	0.44	0.44
7	10	5	0.5	0

for unit 7, where it is stated that the scale elasticity is greater or equal to 1/2. We see that the worry of Golany and Yu (1997) is unwarranted, because for unit 1 we simply have that on the left-hand facet the value of the scale elasticity for unit 1 to the left is infinite, corresponding to the vertical line from unit 1 to the horizontal axis, and on the right-hand facet the value evaluated to the right is 5. Along this facet (the line to unit 2) the scale elasticity is falling, according to the *Regular Ultra Passum Law*

[16] A DEA software developed at the Frisch Centre, Oslo, has been used.

(Frisch, 1965) explored in the DEA setting in Førsund and Hjalmarsson (2002). We see this by the right- and left-hand values of the scale elasticities for unit 2. Since this unit is placed exactly on the line between unit 1 and unit 3 the min- and max values should be equal, and our programme gets this right. We note that the scale elasticity value is less than the right-hand value for unit 1. For unit 3 the left-hand value is less, as predicted, and the right-hand value even smaller. Here we note that the right-hand value is greater than one, and the left-hand value is less than one. This implies that unit 3 is of optimal scale. Unit 4 is placed on the line between unit 3 and 5, and consequently the scale elasticity value is unique. But note that the value is greater than the right-hand value for unit 3. This is in accordance with the finding in Førsund and Hjalmarsson (2002) that the DEA model does not obey the Regular Ultra Passum law for the region of decreasing returns to scale. Unit 6 is also in the line between unit 5 and 7. The unique scale elasticity value is again greater than the right-hand value for unit 5, and also smaller than the left-hand value for unit 7. For the latter the right-hand value is zero, corresponding to a horizontal facet.

In the single output-input case the calculation of the scale elasticity is very simple, since it is the ratio between the marginal- and the average product. Since the marginal product is constant on a facet (line) the scale elasticity has to decrease on a facet where the average product is increasing, as is the case between unit 1 and 3, while the scale elasticity has to increase on facets where the marginal product decreases, as from unit 3 and outwards. The marginal product decreases from facet to facet going outwards from unit 1 by the convexity assumption. On the left-hand side of unit 3 the marginal product is greater than the average, implying increasing returns to scale, while on the right-hand side of unit 3 the marginal product is smaller than the average, implying decreasing returns to scale.

The recurring dataset in Charnes et al. (1994) introduced in Chapter 2 contains both efficient and inefficient observations. Shadow prices on the convexity constraints are also reported. Running our programme yields the results set out in Table 2. The data are illustrated in Figure 1. Units P1, P2, P3 and P4 are located at corners, while units P5, P6 and P7 are inefficient as can be seen in Figure 1. The average scale elasticity is within the interval spanned by the two corresponding scale elasticities. The results for the frontier points C and L confirm the development of the scale elasticity within a

DEA model as analysed in Førsund and Hjalmarsson (2002). The value at C is 2.33, which is less than the lower bound of 3 at P1, and higher than 1.8 which is the upper

Table 2. Scale elasticities and bounds for the Charnes et al. (1994) data

Unit	Input	Output	ε^{in} input-saving	ε^{out} output-increasing	$\bar{e}$ average	ε^{max} upper limit	ε^{min} lower limit
P1	2	2				∞	3.00
P2	3	5				1.80	0.40
P3	6	7				0.57	0.29
P4	9	8				0.38	0.00
P5	5	3	2.33	0.53	0.98		
P6	4	1	∞	0.47	2.50		
P7	10	7	0.57	0.00	0.26		

bound at P2. The value at L is 0.53, which is higher than the lower bound of 0.4 at P2, but lower than the upper bound of 0.57 at P3.

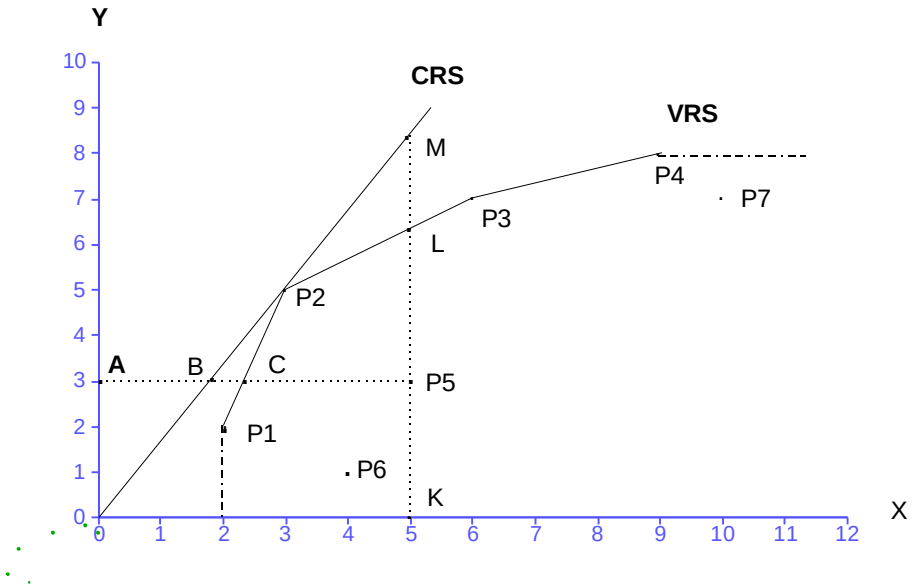

Figure 1. A VRS frontier
Source: Førsund and Hernæs (1994)

The input-corrected frontier point for observation P7 happens to coincide with P3 at a corner. Our LP-solver gives the upper bound as the solution, but this is probably arbitrary, and in such cases both bound values at P3 should be stated. This is an example of the very unlikely occurence with real data of a radial projection ending up at a corner point.

Swedish dairies

Here we have utilised primary data for general milk processing in 28 dairy plants for 1973. The data have been used in Førsund and Hjalmarsson (1979a) and (1979b). Output is measured in tonnes of milk delivered to the plant each year. (The amount of milk received is equal to the amount produced.) The labour input variable is defined as the hours worked by production workers including a technical staff that usually consisted of one engineer. Capital data represents buildings and machines (of user-cost type) and reflects depreciation based on current replacement cost, cost of maintenance and rate of interest.

In the single input case of dairy plants we concentrate on the two scale elasticity values for the radial-projected input- and output points at the frontier for inefficient units and the minimum and maximum values for the efficient ones. Values of infinity are dealt with in a pragmatic way, by truncating the axis at the largest number that it is practical to work with.

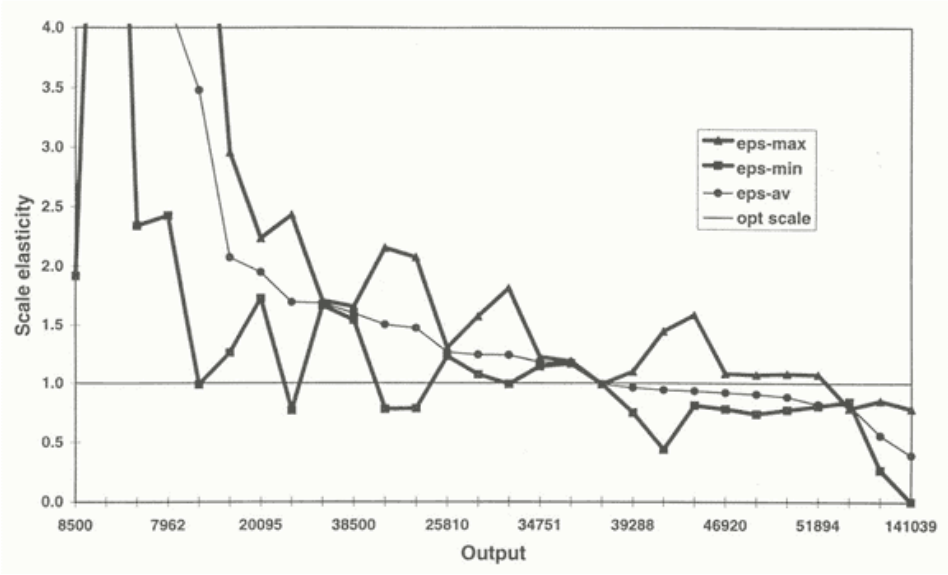

Figure 2: Scale elasticities : Dairy Plants

A picture of the variation in the different scale elasticities is given in Figure 2. The average scale elasticity values plotted for the non-frontier units were calculated according to equation (28), while the mean of min and max values are plotted for the frontier units. The average elasticity for efficient units is calculated as the average of the bounds. The units are sorted according to decreasing values of the average scale elasticity (truncated at 4). The output levels are denoted along the abscissa axis.

Of the 28 dairy plants seven units are on the frontier in the case of variable returns to scale. Of these, three are optimal scale units, i.e. they are frontier units when constant returns to scale is imposed, and while the input-oriented max-value, e^{in} is larger than one, and the output-oriented min-value e^{out} is less than one. The bounds show quite a variation between the input- and output oriented scale elasticity calculations. The more inefficient the unit is the more scope there is for differences. We see that for many units the scale elasticities are on both sides of the strategic value 1. For the three units of optimal scale the lower bound for the first one, starting from the left in the diagram, is 1, while for the two others, located towards the end at the output values of 39288 and 46920, the intervals between the bounds contain the value 1. The four other efficient units are entered with just one point in the diagram.

Except for the three largest units exhibiting decreasing returns to scale at the end of the diagram, there is a lot of variation in output along the descending average scale elasticity curve. Moving along the DEA frontier we have the different observed output levels and the corresponding potential output levels obtained when moving the observed non-frontier units in vertical direction to the frontier, i.e. as the output levels C and L for P5 in Figure 1.

7. Conclusions

The main contribution of the paper is to firmly establish the concept of scale elasticity as defined within the production theory of economics for the DEA model with a piecewise linear frontier. It should be noted that it is not meaningful to ask for the scale elasticity of an inefficient obervation. The scale elasticity may be calculated for

inefficient units by radial projection to the DEA frontier for either input- or output orientation. The formulas for calculation are very simple, being functions of just the relevant efficiency score and the shadow price on the convexity constraint introduced in the DEA formulation to specify variable returns to scale. The procedure is so simple that this route appears the most practical one also for establishing qualitatively whether frontier points exhibit increasing or decreasing returns to scale; confer the discussion started in Banker et al. (1984), Färe et al. (1985), Banker and Thrall (1992), and reviewed in Banker et al. (2000).

The formulas require differentiability. Since efficient observations in general constitute vertices we do not have differentiability at such points, and neither at the reference points (projections including slacks). Following Banker and Thrall (1992) upper and lower bounds on the scale elasticities can be established for such points.

Tone (1996) showed that by extending the solution of the variable returns to scale DEA model to also finding the bounds for the shadow price on the convexity constraint for efficient units, a complete qualitative characterisation of scale properties for all facets can be obtained. This complete characterisation is also obtained as a by-product of using our approach to calculating scale elasticity values.

As for policy conclusions, since projections of inefficient units are hypothetical, the most interesting exercise as to scale properties may be to characterise facets. In addition to using our approach for qualitative characterisation, actually computing the scale elasticity values for inefficient observations give an indication of range of values. The application of our formulas on a real data set showed large variations in scale elaticity values from facet to facet. As demonstrated in Section 3 values may be important for policy analysis regarding the nature of competition that may be sustained. By constructing suitable (inefficient) observations one may check on any scale elasticity value of interest.

References

Afriat, S. (1972): “Efficiency estimation of production functions”. *International Economic Review* 13(3), 568-598.

Aigner, D.J. and S.F. Chu (1968): “On estimating the industry production function”, *American Economic Review* 58, 226-239.

Aigner, D. J. C.A.K. Lovell and P. Schmidt (1977): “Formulation and Estimation of Stochastic Frontier Production Function Models”, *Journal of Econometrics*, 6:1:21-37.

Banker, R. D. (1984): ”Estimating most productive scale size using data envelopment analysis”, *European Journal of Operational Research* 17, 35-44.

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

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

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

Banker, R. D., Charnes, A., and W. W. Cooper (1984): "Some models for estimating technical and scale inefficiencies", *Management Science* 39, 1261-1264.

Banker, R. D. W.W. Cooper, L. M. Seiford, R. M. Thrall and J. Zhu (2000): ”Returns to scale in different DEA models”, forthcoming in *European Journal of Operational Research* (based on a paper presented at the 5th International Conference of the Decision Science Institute, in Athens, Greece, July 4-7, 1999).

Caves, D.W., L.R. Christensen and W. E. Diewert (1982): "The economic theory of index numbers and the measurement of input, output, and productivity", *Econometrica* 50, 1393-1414.

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

Charnes, A., W. W. Cooper, A. Y. Lewin and L. M. Seiford (1994):”Basic DEA models”, Chapter 2 in Charnes, A., W. W. Cooper, A. Y. Lewin and L. M. Seiford (eds.): *Data Envelopment Analysis: Theory, Methodology, and Applications*, Boston/Dordrecht/London: Kluwer Academic Publishers, 23-47.

Cooper, W. W., R. G. Thompson and R. M. Thrall (1996): Introduction: extensions and new development in DEA”, Annals of Operational Research 66, 3-46.

Cooper, W. W., L. M. Seiford and K. Tone (2000): *Data envelopment analysis: a comprehensive text with models, applications, references and DEA-solver software*, Dordrecht/Boston/London: Kluwer Academic publishers.

”Discussion on Mr. Farrell’ paper”, *Journal of the Royal Statistical Society*, Series A (General) 120 (III), 282-290, 1957.

Farrell, M. J. (1957): ”The measurement of productive efficiency", *Journal of the Royal Statistical Society*, Series A (General) 120 (III), 253-281.

Frisch, R. (1965): *Theory of production*, Dordrecht: D. Reidel.

Fukuyama, H. (2000): ”Returns to scale and scale elasticity in data envelopment analysis”, *European Journal of Operational Research* 125, 93-112.

Fukuyama, H. (2001): Returns to scale and scale elasticity in Farrell, Russell and additive models”, *Journal of Productivity Analysis* 16, 225-239.

Färe, R. and S. Grosskopf (1985): ”A nonparametric cost approach to scale efficiency”, *Scandinavian Journal of Economics* 87, 594-604.
Färe, R. and S. Grosskopf (1994): "Estimation of returns to scale using data envelopment analysis: a comment", *European Journal of Operational Research* 79, 379-382.

Färe, R. and C. A. K. Lovell (1978): "Measuring the technial efficiency of production", *Journal of Economic Theory* 19, 150-162.

Färe, R. and D. Primont (1995): *Multi-output production and duality: theory and applications*, Boston/London/Dordrecht: Kluwer Academic Publishers.

Färe, R., S. Grosskopf and C. A. K. Lovell (1983): "The structure of technical efficiency”, *Scandinavian Journal of Economics* 85, 181-190.

Färe, R., Grosskopf. S. and Lovell, C.A.K (1985): *The Measurement of efficiency of production*, Boston: Kluwer,.

Førsund, F.R. (1996): ”On the calculation of the scale elasticity in DEA models”, *Journal of Productivity Analysis* 7, 283-302.

Førsund, F. R. and E. Hernæs (1994): " A comparative analysis of ferry transport in Norway", Chapter 15 in Charnes, A., W. W. Cooper, A. Y. Lewin and L. M. Seiford (eds.): *Data Envelopment Analysis: Theory, Methodology, and Applications*, Boston/Dordrecht/London: Kluwer Academic Publishers, 285-311.

Førsund, F. R. and L. Hjalmarsson (1974): "On the Measurement of Productive Efficiency", *Swedish Journal of Economics*, 76, (2), 141-154.

Førsund, F. R. and L. Hjalmarsson (1979a): Frontier production functions and technical progress: a study of general milk processing in Swedish dairy plants”, *Econometrica*, 47, 883-900.

Førsund, F. R. and L. Hjalmarsson (1979b): "Generalised Farrell measures of efficiency: an application to milk processing in Swedish dairy plants", *Economic Journal*, 89, 294-315.

Førsund, F.R. and L. Hjalmarsson (1987): *Analyses of industrial structure. A Putty-Clay approach*, Stockholm: Almqvist & Wiksell International.

Førsund, F.R. and L. Hjalmarsson (2002): "Are all scales optimal in DEA? Theory and empirical evidence", *Working paper* No 14, ICER.

Førsund, F. R. and Sarafoglous (2002): "On the origins of Data Envelopment Analysis", *Journal of Productivity Analysis* 17, 23-40.

Ganley, J. A. and J. S. Cubbin (1992): *Public sector efficiency measurement. Applications of Data Envelopment Analysis*, Amsterdam/London/New York/Tokyo: North Holland.

Golany, B. and G. Yu (1997): "Estimating returns to scale in DEA", *European Journal of Operational Research* 103, 28-37.
Grosskopf, S. (1986): "The role of the reference technology in measuring productive efficiency", *Economic Journal* 96, 499-513.

Hanoch, G. (1970): "Homotheticity in joint production", *Journal of Economic Theory* 2, 423-426.

Klein, L. R. (1952): *A textbook of econometrics*, Englewood Cliffs, N. J.: Prentice-Hall, Inc., 364-371 (second edition 1974).

Laitinen, K. (1980): *A theory of the multiproduct firm*, Amsterdam-New York-Oxford: North-Holland Publishing Company.

McFadden (eds.): *Production economics: A dual approach to theory and applications*, Vol. 1, Chapter 1, 3-109, Amsterdam: North Holland Publishing Company.

Meeusen, W. and J. van den Broeck (1977): "Efficiency Estimation from Cobb-Dougals Production Functions with Composed Errors", *International Economic Review* 18, 435-444.

Panzar, J. C. and R. D. Willig (1977): "Economies of scale in multi-output production", *Quarterly Journal of Economics* XLI, 481-493.

Seiford, L. M. and J. Zhu (1998): "On alternative optimal solutions in the estimation of returns to scale in DEA", *European Journal of Operational Research* 108, 149-152.

Seiford, L. M. and J. Zhu (1999a): "An investigation of returns to scale in data envelopment analysis", *Omega, International Journal of Management Science* 27, 1-11.

Seiford, L.M. and J. Zhu (1999b): “Sensitivity and stability of the classifications of returns to scale in data envelopment analysis”, *Journal of Productivity Analysis* 12, 55-75.

Starrett, D. A. (1977): "Measuring returns to scale in the aggregate, and the scale effect of public goods", *Econometrica* 45, 1439-1455.

Sueyoshi, T. (1997): “Measuring efficiencies and returns to scale of Nippon Telegraph &Telephone in production and cost analyses”, *Management Science* 43(6), 779-796.

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(11), 1593-1608.

Thanassoulis, E. (2001): *Introduction to the theory and application of Data Envelopment Analysis. A foundation text with integrated software*, Norwell/Dordrecht: Kluwer Academic Publishers.

Thore, S. (1996): “Economies of scale in the US computer industry: an empirical investigation using data envelopment analysis”, *Journal of Evolutionary Economics* 6, 199 – 216.

Tone, K. (1996): “A simple characterization of returns to scale in DEA”, *Journal of the Operations Research Society of Japan* 39(4), 604-613.

Zhu, J. and Z -H. Shen (1995): “A discussion of testing DMUs’ returns to scale”, *European Journal of Operational Research* 81, 590-596.