European Journal of Operational Research 242 (2015) 121–133

Contents lists available at ScienceDirect

European Journal of Operational Research

journal homepage: www.elsevier.com/locate/ejor

Production, Manufacturing and Logistics

Average-cost efficiency and optimal scale sizes in non-parametric analysis☆

Giovanni Cesaroni*, Daniele Giovannola

Department for Local Development, via della Mercede, 9 – 00187 Rome, Italy

ARTICLE INFO

Article history:
Received 6 April 2013
Accepted 29 September 2014
Available online 13 October 2014

Keywords:
Scale efficiency
Returns to scale
Scale economies
FDH
DEA

ABSTRACT

Under fairly general assumptions requiring neither a differentiable frontier nor a constant-returns-to-scale technology, this paper introduces a new definition of an optimal scale size based on the minimization of unit costs. The corresponding measure, average-cost efficiency, combines scale and allocative efficiency, and generalizes the measurement of *scale economies* in efficiency analysis while providing a performance criterion which is stricter than both cost efficiency and scale efficiency measurement. The average-cost efficiency is not reliant upon the uniformity of the firms' input-price vector, and we supply procedures to compute it in both convex and non-convex production technologies. Empirical illustration of the theoretical results is given with reference to large sets of production units.

1. Introduction

A long-established tradition deals with the issue of the optimal scale of production units without assigning a role to input prices. Since the pioneering contribution of Frisch (1965), the optimal scale size has been conceived of as the "technically optimal scale" maximizing average physical productivity along the production function, i.e. the increase in a single output relative to a proportional increase in all inputs. Baumol, Panzar, and Willig (1982) attempted to extend this analysis by including costs and a multiple-output setting, but their ray-average-cost function nevertheless did not account for a change in input proportions as it was only concerned with radial expansions of the given input and output mixes. In this tradition, the optimality of a scale size – that is to say the evaluation of the scale efficiency of a production unit – depends exclusively upon ray average productivity with no role being played by the evaluation of the allocative efficiency of its input mix. By formally developing the concept of most productive scale size (MPSS) for the case of multiple inputs and multiple outputs, Banker's (1984) seminal contribution confirmed this view (see also Färe, Grosskopf, & Lovell, 1985). However, this concept of optimality may become unsatisfactory when one reflects that Farrell (1957) denoted cost efficiency – the ability to minimize total cost borne for a given level of output – as *overall efficiency* precisely because of its capacity to include both the technical and the allocative aspects. Farrell, along with Koopmans (1951) and others, is one of the pioneers of that modern analysis of production which accounts for productive technologies possessing both feasible efficient and inefficient points. The first attempt to integrate scale and cost efficiency into a single measure from this perspective, where applied analysis has extensively relied on non-parametric methods employing linear and integer programming (see, e.g., Banker, Charnes, & Cooper, 1984; Charnes, Cooper, & Rhodes, 1978; Deprins, Simar, & Tulkens, 1984), was made by Färe and Grosskopf (1985). In a data envelopment analysis (DEA) framework, characterized by a convex technology, they introduced the *cost measure of scale efficiency* of a decision making unit (DMU) as the ratio of its constant returns to scale (CRS) cost-efficiency to that of its variable returns to scale (VRS). Sueyoshi (1997, 1999) extended this approach to the measurement of the degree of scale economies.[1] No application of these measures to a non-convex production technology, such that of the free disposal hull (FDH), has yet been found: as an example, De Witte and Marques' (2010) scale measure does not take account of input prices and is therefore not concerned with allocative efficiency.

The kind of integration proposed by Färe and Grosskpof (1985) and Sueyoshi (1999) has, however, proved to be less than effective in prompting the inclusion of allocative efficiency in the definition of an optimal scale size. As a matter of fact, recent contributions dealing with the issue still base their definitions exclusively upon the maximal ray average productivity typical of the MPSS (see, e.g., Førsund & Hjalmarsson, 2004; Podinovski, 2004). In our opinion, one reason

☆ The usual disclaimers apply.

* Corresponding author. Tel.: +39 0667794826; fax: +39 0667792806.
E-mail address: g.cesaroni@governo.it (G. Cesaroni).

[1] The ratio of marginal cost to average cost. In the paper, we will generally speak of 'scale economies' when monetary costs are being considered. Reference to productivity in physical terms (input-output) will be denoted by the use of 'returns to scale'.

http://dx.doi.org/10.1016/j.ejor.2014.09.062
0377-2217/

List of abbreviations

ACE	average-cost efficiency
CRS	constant returns to scale
DEA	data envelopment analysis
DMU	decision making unit
FDH	free disposal hull
MPSS	most productive scale size
OE	overall efficiency/cost efficiency
OSS	optimal scale size
TSRE	technical and scale radial efficiency
VRS	variable returns to scale

for this outcome might lie in the fact that Färe and Grosskopf's (1985) methodology does not envision a separation between scale and allocative efficiencies. As a consequence, their framework is unfit: a) to ascertain the logical relationship existing between primal (production-based) and dual (cost-based) optimal scale sizes[2]; b) to establish a clear-cut algebraic relationship between production-based and cost-based scale efficiencies.

Our analysis aims at offering a solution to these research issues – including the introduction of the cost measure of scale efficiency in FDH technologies – under assumptions more general than those featured in the methodology of Färe and Grosskopf (1985) and Sueyoshi (1999). In fact, only free disposal and VRS are required, thereby relaxing the assumptions of convexity, CRS and differentiability of the production frontier. The purpose is that of integrating the research work initiated by Banker (1984) – characterized by VRS, MPSS and a simple method for the determination of returns to scale – with Färe and Grosskopf's (1985) cost measure of scale efficiency.

To pursue this objective, we have introduced a new definition of optimal scale size (OSS) as the scale which minimizes the ray average cost, taking into account the allocative efficiency of the input mix. In particular, we propose an efficiency measure based upon the ratio between the ray average cost of a DMU's output vector – evaluated at its OSS – and its total cost. The measure is called average-cost efficiency because it can be interpreted as the potential reduction in unit cost that a DMU could achieve if it adopted the scale of production and the input mix of its OSS. This efficiency criterion is shown to be more restrictive than both VRS traditional measures of cost efficiency and scale efficiency based on ray average productivity.

The outline of the paper is as follows. Under the free disposal assumption, Section 2 introduces the definitions of average-cost efficiency (ACE) and of OSS. Without having recourse to the usual regularity conditions, Section 3 characterizes an OSS and its relationship with an MPSS by means of the decomposition of the ACE measure into the sum of a scale and an allocative component. In general, an OSS need not coincide with an MPSS, precisely because of the potential allocative inefficiency of the latter. Section 4 is dedicated to illustrating the geometric and economic significance of our measure, and to its computation in convex and non-convex technologies, the last sub-section clarifying the role of our assumptions on input prices. Section 5 discusses the relationship of our analysis with the literature on scale and cost-scale efficiency, while pointing out the advantages that ACE measure achieves over the existing approaches. Detailed empirical applications, regarding the convex and non-convex case, are presented in Section 6, while Section 7 offers some concluding remarks on the features of our proposed performance-criterion and the potential for application in various economic sectors.

[2] We use primal and dual with the same meaning assigned them by Färe and Grosskopf (1985); see ibid., p. 595. The same use of these words is made by Sueyoshi (1999) and Tone and Sahoo (2006).

2. Definition of average-cost efficiency and of an optimal scale size

The production technology upon which the analysis is based satisfies only the free disposability assumption, as it is the FDH of the observed production possibilities. Free disposal means that, given an observed input–output vector which we assume to belong to the technology, we postulate as a feasible point of the same technology any other vector which: a) has the same outputs, and inputs which are no smaller, b) has the same inputs, and outputs which are no larger. In other words, excess inputs and excess outputs can be disposed of at no cost. We adopt the original FDH model (e.g. Tulkens, 1993) in so far as it makes no specific assumption on the returns-to-scale regime, which is therefore VRS. Introducing some notation, we have n DMUs, with each DMU j $(j = 1, \ldots, n)$ using m inputs, $x_{ij} (i = 1, \ldots, m)$, to produce s outputs, $y_{rj} (r = 1, \ldots, s)$. We assume that the observed input and output vectors of a generic DMU are, respectively, $\mathbf{x}_j = (x_{1j}, \ldots, x_{mj})' \geq \mathbf{0}$ and $\mathbf{y}_j = (y_{1j}, \ldots, y_{sj})' \geq \mathbf{0}$, where the prime indicates the transposition operation.[3] If we denote the $m \times n$ matrix of inputs as $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_n]$ and the $s \times n$ matrix of outputs as $\mathbf{Y} = [\mathbf{y}_1, \ldots, \mathbf{y}_n]$, then the production possibility set is expressed as

$$T = \left\{ (\mathbf{x}, \mathbf{y}) | \mathbf{X}\lambda \leq \mathbf{x}, \mathbf{Y}\lambda \geq \mathbf{y}, \sum_{j=1}^{n} \lambda_j = 1, \lambda_j \in \{0, 1\};\ j \in J \right\} \quad (1)$$

where λ is the $n \times 1$ vector with components equal to λ_j, and $J = \{1, \ldots, n\}$.

Being dependent exclusively upon the strong disposability assumption, the conclusions we reach carry over to a convex VRS production possibility set, i.e. a set which is obtained by simply changing the constraint on λ_j in (1), and which is defined as

$$T^{DEA} = \left\{ (\mathbf{x}, \mathbf{y}) | \mathbf{X}\lambda \leq \mathbf{x}, \mathbf{Y}\lambda \geq \mathbf{y}, \sum_{j=1}^{n} \lambda_j = 1, \lambda_j \geq 0;\ j \in J \right\} \quad (2)$$

For future reference, we note that the CRS versions of the above technologies, which we denote respectively as T_{CRS} and T_{CRS}^{DEA}, are obtained from (1) and (2) by substituting λ with $\mathbf{z}$, a vector whose j-th element is $z_j = w\lambda_j$ – where $w > 0$ is an arbitrary scaling factor.

As far as costs are concerned, we posit that each DMU faces the same vector of input prices, $\mathbf{p} = (p_1, \ldots, p_m) > \mathbf{0}$, $\mathbf{p}\mathbf{x}_j$ thus being the actual total cost that DMU j bears for producing its output vector, $\mathbf{y}_j$. Uniformity of input prices across the different DMUs is being assumed only for notational convenience, while most of our results will in fact stand even when DMUs face varying input-price vectors (see, below, Section 4.3).

In the literature on efficiency analysis, we have not found an efficiency concept involving the ratio of the average cost of a DMU's output to that of a reference unit. In fact, as previously stated, the measures of both Färe and Grosskopf (1985) and Sueyoshi (1997), while combining cost and scale efficiency, refer only to the total cost borne by a DMU for producing its output vector. A ratio relating to unit costs can however be devised in connection with the concept of ray average cost. Hence, we will consider the ratio between the ray average cost evaluated at a reference unit and that evaluated at the DMU's current scale size (i.e. total cost).

Following Baumol's basic definition (see Baumol, 1977, p. 811; Baumol et al., 1982, pp. 48–49), the ray average cost of the generic DMU j is given as

$$RAC(\mathbf{y}_j) = \frac{C(t\mathbf{y}_j)}{t} \quad (3)$$

[3] The vector-inequality sign means that each element of the left vector is weakly greater than the corresponding element of the right vector, with at least one component being strictly greater.