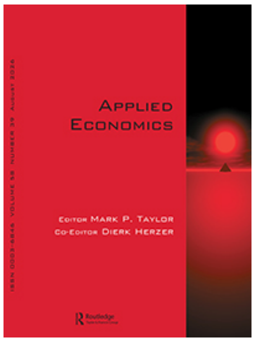

Applied Economics

ISSN: 0003-6846 (Print) 1466-4283 (Online) Journal homepage: www.tandfonline.com/journals/raec20

The shape of ray average cost and its role in multioutput scale economies: some generalizations

Giovanni Cesaroni, Kristiaan Kerstens, Jafar Sadeghi & Ignace Van de Woestyne

APPLIED ECONOMICS
2026, VOL. 58, NO. 39, 8179–8194
https://doi.org/10.1080/00036846.2025.2536874

The shape of ray average cost and its role in multioutput scale economies: some generalizations

Giovanni Cesaroni[a], Kristiaan Kerstens[b], Jafar Sadeghi[c] and Ignace Van de Woestyne[d]

[a]Universitas Mercatorum, Rome, Italy; [b]Univ.of Lille, CNRS, IESEG School of Management, UMR 9221 - LEM - Lille Économie Management, Lille, France; [c]Royal Military College of Canada, Kingston, Canada; [d]KU Leuven, Research Centre for Operations Research and Statistics (ORSTAT), Brussels, Belgium

ABSTRACT
Establishing a link between the so-called 'neoclassical' and 'axiomatic' approaches to production theory, we deal with some central and unresolved issues concerning scale economies in multi-output technologies. First, we extend Panzar and Willig's (1977) results on the duality between primal and dual scale elasticity measures, which helps pointing out the unknown role played in this regard by the monotonicity of the local degree of homogeneity of the technology. Second, under a general representation of a convex technology – allowing for non-differentiability of the cost function and multiple optima – we determine the shape of the ray average cost function. Third, in the same setting, we determine an unambiguous relationship between cost scale elasticity and cost scale efficiency, and therefore between local and global scale economies. Fourth, we develop a complete map of values taken by primal and dual scale elasticities and point out that the equality between returns to scale and scale economies local measures breaks down in a convex technology at points where the cost function is not differentiable. This highlights the importance of considering non-differentiability when analysing scale economies in production technologies.

KEYWORDS
Data envelopment analysis; returns to scale; scale economies; ray average cost

JEL CLASSIFICATION
C61; D24; L25

I. Introduction

The ray average cost (RAC) extends the traditional concept of average cost, frequently used in economics, from the case of the production function to the case of multi-output production technologies. Intuitively, for a given output mix and its associated cost of production, the RAC shows how this cost changes with a proportional increase/decrease of the scale of the output mix, thus offering information which is similar to that of the average cost considered in the traditional theory of the firm (Cassels 1937).

The introduction of RAC as a tool to investigate scale economies in multiple output technologies dates back to the seminal contributions of Baumol (1977), Panzar and Willig (1977) and Baumol, Panzar, and Willig (1982). Following Färe, Grosskopf, and Lovell (1988, 721–722), these contributions fall within the 'neoclassical approach' to scale economies measurement, which is broadly characterized by the use of the transformation and cost functions, and the recourse to quantitative and local measures of scale economies such as the primal (i.e. production) and dual (i.e. cost) scale elasticities – which are defined in a neighbourhood of a given scale of activity. In the competing 'axiomatic' approach initiated by Koopmans (1957), the nature of scale economies is 'deduced from the structure of the production set, or from a comparison of the production set with larger production sets exhibiting known scale behavior' (Färe, Grosskopf, and Lovell 1988, 721), with the comparison yielding qualitative information on the nature of scale economies (see, e.g. Banker, Charnes, and Cooper 1984; Färe, Grosskopf, and Lovell 1983), i.e. the direction towards an efficient scale size. This approach is based on the computation of scale efficiency measures and it generally disregards the local behaviour of the cost function, even in the only instance when the analysis of costs is considered (Färe and Grosskopf 1985). Thus, while the 'neoclassical' approach focuses on local scale economies in terms of both production and cost analyses, the 'axiomatic' approach is mainly

CONTACT Kristiaan Kerstens k.kerstens@ieseg.fr Univ.of Lille, CNRS, IESEG School of Management, UMR 9221 - LEM - Lille Économie Management, Lille F-59000, France PH: +33 320545892
Supplemental data for this article can be accessed online at https://doi.org/10.1080/00036846.2025.2536874

concerned with global[1] measures in returns to scale analysis in production, such as the scale efficiency determined by Banker's (1984) notion of a most productive scale size (MPSS). This author proves that at this MPSS – which achieves a global maximum for average productivity and is scale efficient – the primal scale elasticity equals one, but he does not address the issue of the relationship between this scale elasticity and scale efficiency at points that are not an MPSS.[2]

A first attempt to establish a direct link between scale elasticity and scale efficiency measures over the production technology, as a function of the scaling factor, is offered in Färe, Grosskopf, and Lovell (1988). However, their results are limited to the analysis of returns to scale – characterized by proportional changes in inputs and outputs – in the specific case in which the regular ultra-passum law holds (see Frisch 1965). Moreover, these authors do not consider the fact that the definition of the degree of scale economies taken from Panzar and Willig (1977) is not valid for any behaviour of the local degree of homogeneity along the production frontier, a problem that the literature has not yet addressed.

A second important contribution for the integration of both approaches is Chavas and Cox (1999, 308) who are the first to explicitly define the RAC function in the context of 'axiomatic' technologies and to point out that its minimum over this set is strictly connected to the evaluation of allocative inefficiency *via* the constant returns to scale (CRS) cost function, which implies non-equiproportionate changes in inputs (see also Proposition 1 in Baumol 1977, 812). However, the impossibility of establishing the behaviour of the RAC function in a general technology – an unsolved issue in the neoclassical approach (see, e.g. Baumol 1977; Baumol and Fischer 1978; Chambers 1988) – has led these authors to conclude that their proposed cost scale efficiency measure is a criterion to detect the presence of local CRS but not a criterion of global CRS which instead is concerned with the direction towards an optimal scale size (OSS), the scale size that determines the global minimum of the RAC.[3]

The recourse to cost analysis and the RAC concept is required by the necessity of taking into account the change in input proportions associated with scale effects. This element is absent in the more traditional analysis of the axiomatic approach based on returns to scale (RTS), where only the ratio between the maximum proportional change in outputs to that in the inputs along the frontier of a variable returns to scale (VRS) technology – i.e. the ray average productivity (RAP) – is considered (see Banker and Thrall 1992; Banker, Chang, and Cooper 1996, section 2.1 in; Tone and Sahoo 2003), while it is also overlooked in the comprehensive analysis of Balk (2001). In fact, Cesaroni and Giovannola (2015) show that in a general VRS technology, scale efficiency measures, based, respectively, on RAP and RAC, determine at an efficient point different global classifications from both the quantitative and the qualitative point of view. This is indirectly confirmed by Zelenyuk (2014), who proves that – in a general technology – homotheticity is a sufficient condition for the equivalence of the two measures at a cost-efficient point. This clarifies why Podinovski (2004, 2017) results on the monotonic behaviour of the RAP function in a convex technology – which complement the Banker (1984) and the Banker and Thrall (1992) analyses of the relationship between scale efficiency and primal scale elasticity – may not apply to RAC. As a consequence, the coincidence between local and global RTS indicators may not necessarily hold for scale economies (SE) indicators.

The global behaviour of RAC and cost functions is important not only to nonparametric models of efficiency analysis in multi-output and input technologies such as the Data Envelopment Analysis (DEA) model of (Banker, Charnes, and Cooper 1984) but also to parametric models based on the estimation of input distance and cost functions – relying on the duality between primal and dual scale elasticities (see, e.g. Berger 1993; Coelli and Perlman 1999; Ferrier and Lovell 1990).

For these reasons, our work pursues a better integration of the 'axiomatic' and 'neoclassical'

[1]The adjective global for returns to scale is introduced in Podinovski (2004), with the same meaning it has in mathematical analysis when referring to the maximum/minimum of a function over its whole domain.

[2]Corollary 1 in Banker (1984, 40) employs a global definition of returns to scale based on scale efficiency, which is different from the local one based on scale elasticity (see, ibid., expression 2, p. 36).

[3]The divergence between global and local indicators of scale economies in a general technology is discussed in Appendices 4 and 5.2.

approaches by investigating SE in terms of RAC in a convex technology which allows for multiple optima and for a cost function that is not-everywhere differentiable. In particular, the properties of the cost function permit to clarify the above unresolved issues: (i) the determination of the shape of the RAC, (ii) the relation between cost scale efficiency and dual scale elasticity measures as a function of the scaling factor, and (iii) the correspondence between primal and dual scale elasticity measures when differentiability of the cost function does not hold. As a by-product of these clarifications, we also obtain a general model for the determination of the global SE.

The main research gaps addressed by our work are visualized in Table 1. In Table 1, the grey shaded area differentiates the convex technology from the nonconvex one, while the main research problems for which we offer a solution are denoted by the comments in italics.

This contribution is structured as follows. Section II introduces assumptions and some preliminary definitions, along with a correction of the duality relationship between primal and dual scale elasticities of Panzar and Willig (1977). Section III derives the shape of RAC in the presence of multiple optima on the basis of a Shephard's (1970) property of the cost function in the output space. Section IV shows the ensuing relationships between cost scale efficiency and primal and dual scale elasticities. Section V applies some results of the previous section to determine the exact relation between RTS and SE classifications in a convex technology. Section VI illustrates the main results by means of an empirical application based on two secondary data sets using a nonparametric multiple input and output technology. Finally, Section VII summarizes our contributions.

II. Production technology and duality between scale elasticity measures

Definitions and review

The multiple input and multiple output production processes are described as an m input vector $\boldsymbol{x} \in X$ used in the production of an s output vector $\boldsymbol{y} \in \boldsymbol{Y}$, where $\boldsymbol{X} \subset \boldsymbol{R}_{+}^{m}$ and $\boldsymbol{Y} \subset \boldsymbol{R}_{+}^{s}$ are compact sets. The production possibility set or technology is represented by the feasible set:

$$T = \left\{(\boldsymbol{x}, \boldsymbol{y}) \in \boldsymbol{R}_{+}^{m+s} | \boldsymbol{y} \text{ can be produced from } \mathbf{x}\right\} \tag{1}$$

where T is a closed subset of $\boldsymbol{X} \times \boldsymbol{Y}$ which we assume to satisfy some regularity conditions[4] (see, e.g. Diewert 1973; Panzar and Willig 1977):

(A.1) $(0, \boldsymbol{y}) \notin T$ unless $\boldsymbol{y} = 0$;

(A.2) T is non-empty and there exists $\boldsymbol{x} \neq 0$ and $\boldsymbol{y} \neq 0$ such that $(\boldsymbol{x}, \boldsymbol{y}) \in T$;

(A.3) for $(\boldsymbol{x}, \boldsymbol{y}) \in T$ and $(\boldsymbol{x}', \boldsymbol{y}') \in \boldsymbol{X} \times \boldsymbol{Y}$, $\boldsymbol{x}' \geq \boldsymbol{x}, \boldsymbol{y}' \leq \boldsymbol{y}$ implies $(\boldsymbol{x}', \boldsymbol{y}') \in T$;

(A.4) if $(\boldsymbol{x}, \boldsymbol{y})$ and $(\boldsymbol{x}', \boldsymbol{y}') \in T, 0 \leq \lambda \leq 1$implies that $\lambda(\boldsymbol{x}, \boldsymbol{y}) + (1 - \lambda)(\boldsymbol{x}', \boldsymbol{y}') \in T$;

(A.5) the transformation function $\Phi(\boldsymbol{x}, \boldsymbol{y})$ is piecewise continuously differentiable.

The above conditions state that in the feasible set (i) no free lunch is allowed and $(0, 0)$ is included,

Table 1. Research gaps in the literature.

	Global and local RTS	Global and local SE	Global RTS and SE (non-homotheticity)	Local RTS and SE (differentiability)	Local RTS and SE (non-differentiability)	RAC shape
Convex	Banker (1984) Banker and Thrall (1992) Podinovski (2004) Podinovski (2017)	*No result*	Zelenyuk (2014) Cesaroni and Giovannola (2015)	Panzar and Willig (1977); Färe and Primont (1995)	Panzar and Willig (1977); Sueyoshi (1999) *Partially correct results*	Färe (1974) for single-product case *No result for multiproduct case*
Nonconvex	Podinovski (2004) Cesaroni, Kerstens, and Van de Woestyne (2017)	*No discussion of a smooth technology*	Zelenyuk (2014) Cesaroni and Giovannola (2015)	Panzar and Willig (1977); Färe and Primont (1995)	Panzar and Willig (1977) *Incorrect result*	Baumol, Panzar, and Willig (1982) No specific Shape

[4]In the following, the notation for vector equalities/inequalities means that a certain relationship holds for each of the corresponding components of the two vectors considered.

(ii) there are some non-trivial production processes, i.e. processes featuring at least one positive input and one positive output, (iii) inputs and outputs are freely (strongly) disposable within the bounds of $\boldsymbol{X} \times \boldsymbol{Y}$, and inaction is possible (i.e. $\boldsymbol{y} = 0$), (iv) the feasible set is convex, (v) continuous differentiability of $\Phi(.)$ does not hold everywhere. Note that compactness of T and conditions (A.1–A.3) are also assumed in Baumol, Panzar, and Willig (1982), 52).

According to Panzar and Willig (1977) and McFadden (1978), compactness of T and conditions (A.1–A.3) are sufficient to establish that an efficient production exists and that a continuous production transformation function,[5]

$$\Phi(\boldsymbol{x}, \boldsymbol{y}) \geq 0 \Leftrightarrow (\boldsymbol{x}, \boldsymbol{y}) \in T, \tag{2}$$

with $\frac{\partial \Phi}{\partial x_h} \geq 0$ and $\frac{\partial \Phi}{\partial y_k} \leq 0$ exists – where h and k denote a generic input and output, respectively, and $\Phi(\boldsymbol{x}, \boldsymbol{y}) = 0$ at an efficient point. These same conditions are also necessary and sufficient for the existence of a cost function C that is positive for positive outputs and weakly increasing in prices w_h and outputs y_k, where

$$C(\boldsymbol{y}, \boldsymbol{w}) = min\{\boldsymbol{wx} | (\boldsymbol{x}, \boldsymbol{y}) \in T\}, \tag{3}$$

with $\boldsymbol{w} \in \boldsymbol{R}^m_{++}$ a vector of input prices (see McFadden 1978). Moreover, compactness and convexity of T ensure that C is continuous in outputs (see Appendix 2, proof of Theorem 2). The argmin of expression (3), i.e. an input vector which is cost-efficient for $\boldsymbol{y}$ at $\boldsymbol{w}$, is denoted as $\boldsymbol{x}^*(\boldsymbol{w}, \boldsymbol{y})$.

In the remainder of the contribution, an assumption alternative to (A.1) is also considered:

(A.1bis) $(0, \boldsymbol{y}) \notin T$ with $\boldsymbol{y} \geq 0$.

This assumption excludes $(0, 0)$ from T, but it does not affect the existence and the properties of $\Phi(.)$ (see McFadden 1978, section VII).

As far as RTS is concerned, we observe that no assumption is imposed on T, which in principle is a VRS technology. For future reference, we remark that the CRS extension of technology T is defined as $T^{CRS} = \{(\boldsymbol{x}, \boldsymbol{y}) : (\lambda \boldsymbol{x}, \lambda \boldsymbol{y}) \in T \text{for some} \lambda > 0\}$.[6]

Duality relationships

Following Panzar and Willig (1977), we define production scale elasticities in multi-output production as follows:

$$\mathrm{S} \equiv -\frac{\sum_h x_h \frac{\partial \Phi}{\partial x_h}}{\sum_k y_k \frac{\partial \Phi}{\partial y_k}} \tag{4}$$

Here, S is the primal scale elasticity, measuring the ratio between the proportional increase in outputs to the proportional increase in inputs in a neighbourhood of an efficient point $(\boldsymbol{x}, \boldsymbol{y})$. This ratio provides insight into returns to scale from the production (primal) perspective. Following Panzar and Willig (1977), we define cost scale elasticities as:

$$\hat{\mathrm{S}} \equiv \frac{C(y, w)}{\sum_k y_k C_k(y, w)} \tag{5}$$

In this expression, $\hat{\mathrm{S}}$ is the dual scale elasticity, indicating the ratio between the proportional increase in total cost to the proportional increase in outputs in a neighbourhood of an efficient point. This can be interpreted as the ratio of average to marginal cost, which captures the scale economies from the cost (dual) perspective, indicating whether locally increasing production leads to a lower or higher average cost.

To reformulate[7] corollary 2 in Panzar and Willig (1977, 491), which is valid for both convex and nonconvex technologies, at an efficient point $(\boldsymbol{x}, \boldsymbol{y})$, we determine $r(\lambda)$ – the local degree of homogeneity of the technology – as:

$$\Phi\left(\lambda \boldsymbol{x}, \lambda^{r(\lambda)} \boldsymbol{y}\right) = 0, \tag{6}$$

and define

$$r \equiv lim_{\lambda \to 1} r(\lambda). \tag{7}$$

where λ is a positive scalar.

We are now in a position to establish the desired result:

Proposition 1: Under either assumptions (A.1)-(A.3) or (A1bis)-(A.3), and (A.5), at a cost-efficient point $(\boldsymbol{x}^*(\boldsymbol{w}, \boldsymbol{y}), \boldsymbol{y}) \in T$ where $\Phi(.)$ is continuously

[5]See McFadden (1978, 26). Analogous conditions for the existence of a transformation function are in Diewert (1973) for the convex case.
[6]Observe that the CRS technology includes (0,0) if and only if this point belongs to the VRS technology.
[7]A specific inconsistency affects Panzar and Willig's result, as shown in Appendix 1. The same problem applies to propositions 3C1 and 3C3 in Baumol, Panzar, and Willig (1982, 55–57).

differentiable, the following equalities hold $r = S = \hat{S}$.

Proof: All proofs of the results are found in Appendix 2.

An analogous proposition is offered in Färe and Primont (1995), 53; (see also Tone and Sahoo 2003, 173), but our proof differs because of the use of the transformation function and of the ensuing possibility of establishing an immediate link between scale elasticities, RAP and the local degree of homogeneity r of a technology (see Appendix 1).

Proposition 1bis: Under either assumptions (A.1)-(A.3) or (A.1bis)-(A.3), and (A.5), at a cost-efficient point $(\boldsymbol{x}^*(\boldsymbol{w},\boldsymbol{y}),\boldsymbol{y}) \in T$ where $\Phi(.)$ is not continuously differentiable, S and $\hat{S}$ can be in any relationship.

This original result follows from the invalidity of proposition 5 in Panzar and Willig (1977, 491). It clarifies that discontinuities prevent from establishing any specific relationship between primal and dual scale elasticities in the case of a nonconvex technology.

III. Ray average cost

The concept of RAC is a tool to investigate scale economies in multiple output technologies and dates back to the seminal contributions of Baumol (1977), Baumol and Fischer (1978) and Baumol, Panzar, and Willig (1982) in a setting where eventual inefficiency in technology is ignored. By contrast, Panzar and Willig (1977) allow for inefficiency in the production technology, but – differently from the preceding authors – overlook the issue of the shape of the RAC function, as a consequence of their specific focus on the relationship between primal and dual measures of scale elasticity at a point of the cost function. Chavas and Cox (1999, 308) are the first to explicitly define the RAC function in the context of VRS technologies with inefficiency and to point out that this concept is strictly connected to the evaluation of scale inefficiency with respect to the CRS cost function, i.e. the cost function (3) but then defined with regard to T^{CRS}.

Contrary to the more traditional approach in economics based on average productivity, this strand determines SE by means of the behaviour of the cost function in response to equiproportional variations in outputs at given input prices (see, e.g. Baumol 1977, 811 and Proposition 1 (p. 812)). Constant SE arises when the average cost remains stationary at a minimum level. Otherwise, we have an increasing or decreasing SE depending on whether the average cost is decreasing with an increase or with a decrease in the output's scale size, respectively (see definition 5 in Panzar and Willig 1977; see also Tone and Sahoo 2003). In any case, to the best of our knowledge no contribution addresses the issue of the shape of the RAC in a multiple output setting, which has so far been conceived as an assumption rather than a consequence of the form of the technology (see, e.g. Baumol, Panzar, and Willig 1982, 257; Chavas and Cox 1999, 307).

In the following, we address this issue under general conditions in the sense that we allow both for multiple minima of the RAC and for non-differentiability of C.

Preliminaries: RAC function, optimal scale sizes, scale efficiency, multiple minima

Following Baumol (1977, 811) and Baumol, Panzar, and Willig (1982, 48–49), we define the RAC of a given output vector $\boldsymbol{y}_j \neq 0$ as a function

$$RAC_{yj} : \boldsymbol{R}_+ \mapsto \boldsymbol{R}_+ : t \mapsto RAC_{yj}(t) = RAC\left(t, \boldsymbol{y}_j\right) \text{ with}$$

$$RAC\left(t, \boldsymbol{y}_j\right) = \frac{C\left(t\boldsymbol{y}_j\right)}{t} \quad t > 0, \tag{8}$$

where for notational convenience we have suppressed the input price vector $\boldsymbol{w}$ – which we assume as given – from the cost function in the numerator and $C\left(t\boldsymbol{y}_j\right) = \boldsymbol{w}\boldsymbol{x}^*$ with $\left(\boldsymbol{x}^*\left(\boldsymbol{w}, t\boldsymbol{y}_j\right), t\boldsymbol{y}_j\right) \in T$ and t is a positive scalar.

The RAC function RAC(t,$\boldsymbol{y}_j$) represents the average cost per unit output when the output vector $\boldsymbol{y}_j$ is scaled by t. RAC allows us to analyse how average costs change when we scale production up or down along the ray defined by $\boldsymbol{y}_j$: it provides insights into scale economies. This can be easily understood by means of Figure 1, where a U-shaped RAC of $\boldsymbol{y}_j$ is depicted as a function of the scaling factor t, which is displayed on the horizontal axis, and where constant SE holds in the interval $[t_1, t_2]$.

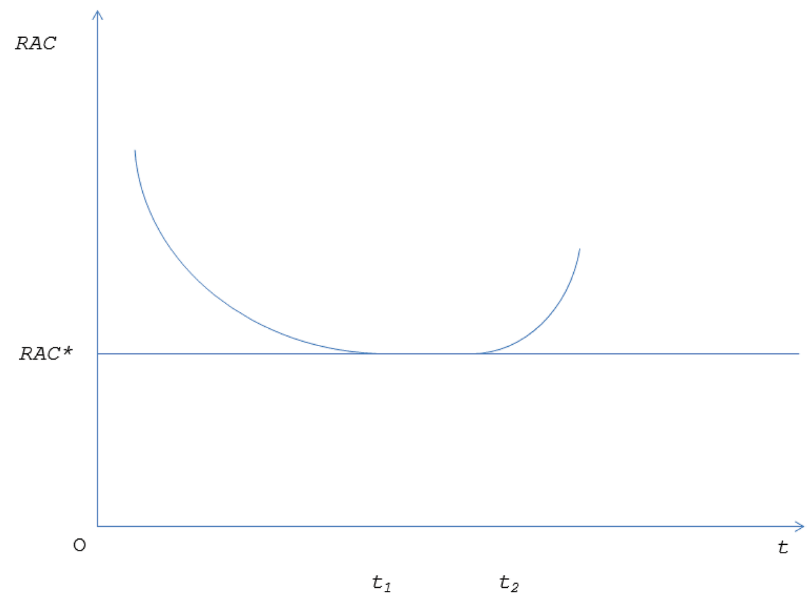

Figure 1. U-shape of RAC with multiple minima.

To define the RAC function (8) also in $t = 0$, we employ the limit definition

$$RAC\left(0, \boldsymbol{y}_j\right) = \lim_{t \to 0} \frac{C\left(t\boldsymbol{y}_j\right)}{t} \tag{9}$$

Chavas and Cox (1999, 307–308) establish that whenever a positive minimum of (8) with respect to t ($RAC^* > 0$) exists in T, this equals the minimum cost of $\boldsymbol{y}_j$ in T^{CRS}. In other words, at given input prices the minimum RAC of an output mix $\boldsymbol{y}_j \neq 0$ in a VRS technology corresponds to the value determined by the CRS cost function. Following Cesaroni and Giovannola (2015, 123), a radial scaling factor between the output vectors of two different production possibilities belonging to T – the one under examination $\left(\boldsymbol{x}_j, \boldsymbol{y}_j\right)$ and a generic reference unit $\left(\boldsymbol{x}_i, \boldsymbol{y}_i\right)$ – can be derived from this operational definition[8]

$$\gamma_{j,i} = \max_k \left\{\frac{y_{kj}}{y_{ki}}\right\}, \; where \; \gamma_{j,i} \in (0, \infty] \tag{10}$$

where y_{kj} is the k-th output of unit j, and k = 1, . . . ,s. Similarly, y_{ki} denotes the k-th output of a generic reference unit i. Here, $\gamma_{j,i}$ represents the maximum proportion by which the outputs of unit i need to be scaled up (or down) to at least match the outputs of unit j in every output dimension. This scaling factor ensures that for all outputs $y_{kj} \leq \gamma_{j,i}\, y_{ki}$.

By evaluating RAC_{yj} (t) at $t = \frac{1}{\gamma_{j,o}}$ for all $\left(\boldsymbol{x}_i, \boldsymbol{y}_i\right) \in T$, we can identify the unit that minimizes the RAC function along the ray through $\boldsymbol{y}_j$. We denote the index of this unit as ‘o’, representing the OSS for $\boldsymbol{y}_j$. Therefore, the minimizing value of RAC_{yj} (t) occurs at $\frac{1}{\gamma_{j,o}}$, where $\gamma_{j,o}$ is the scaling factor associated with the OSS. The OSS corresponds to the unit $\left(\boldsymbol{x}_o, \boldsymbol{y}_o\right) \in T$ where $\boldsymbol{y}_o = \frac{1}{\gamma_{j,o}} \boldsymbol{y}_j$. Therefore, if we define an efficiency measure as the ratio

$$R_j^* = \frac{RAC^*}{C\left(\boldsymbol{y}_j\right)}, \tag{11}$$

then it is evident that this efficiency measure satisfies $R_j^* \leq 1$. The efficiency measure equals one if and only if $\left(\boldsymbol{x}_j, \boldsymbol{y}_j\right)$ is an OSS (i.e. if $\gamma_{j,o} = 1$).[9] Otherwise, it is less than one. Moreover, R_j^* can be readily interpreted as the cost measure of scale efficiency of point

[8]When 0/0 occurs for a component, this is excluded from the comparison; moreover, we consider the set of affinely extended real numbers. Therefore, we remark that, for a positive output vector under examination, (9) allows for the consideration of a reference unit with a zero output vector (i.e. gamma equals infinity).

[9]See Proposition 3 in Cesaroni and Giovannola (2015, 124), which states that ratio (10) equals unity for the output mix of an OSS. Observe that this ratio can also be assumed to equal unity in cases where RAC is minimized at (0,0): this allows to consider (0,0) as an OSS.

$\left(x^{*}\left(w_j, y_j\right), y_j\right)$[10] (see Färe and Grosskopf 1985). Furthermore, the associated $\gamma_{j,o}$ is a straightforward indicator of global SE, that is the direction in T along which outputs can be moved to achieve the minimum RAC. Note that this indicator, besides supplying the qualitative information on SE, also gives quantitative information about the actual distance between the output mix y_j and that of the OSS. The concepts introduced in definitions (10) and (11) are illustrated by means of a numerical example in Appendix 3.

Finally, as far as the multiplicity of OSS points with $RAC^{*} > 0$ is concerned, we present the following useful lemma.

Lemma 1: Under either assumptions (A.1–A.3) or (A1.bis)-(A.3), the OSS of a cost-efficient point $\left(x_j^{*}, y_j\right) \in T$ with a finite $\gamma_{j,o}$ and $RAC^{*} > 0$ is unique up to a positive multiplicative constant.

Lemma 1 reveals that multiple solutions to the minimization of RAC can occur if and only if exact proportional replicas of an OSS (x_o, y_o) are present in the technology.

RAC function in a nonconvex technology

Given the complexity of the topic pointed out by Proposition 1bis, this section is only meant to briefly illustrate the behaviour of the RAC function in a technology satisfying the conditions of Proposition 1. In this setting, as clearly shown in Baumol and Fisher (1978, section VI), the cost function C is not convex in outputs. Therefore, the case of multiple minima discussed in Lemma 1 can be illustrated as in Figure 2, where the shape of the VRS cost function in a nonconvex technology is considered. Note that, for given input prices, the RAC function is given in this figure as the ratio $C\left(ty_j\right)/t$, that is as the slope of the ray joining the origin and a point on $C\left(ty_j\right)$.

This Figure 2 helps to understand that by using the cost scale efficiency criterion for global SE based on $\gamma_{j,o}$ we have four possible cases: (i) constant SE (points A and B), (ii) increasing SE ($\gamma_{j,o} < 1$ points situated at the left of A), (iii) decreasing SE ($\gamma_{j,o} > 1$ points positioned at the right of B) and (iv) sub-constant SE (both $\gamma_{j,o} < 1$ and $\gamma_{j,o'} > 1$ points situated between A and B) (see Cesaroni, Kerstens, and Van de Woestyne 2017, 1446).

The case of a single RAC minimum is shown in Figure 3. Here, we may point out that the cost scale elasticity criterion for local SE gives misleading information regarding global SE as determined by cost scale efficiency: in fact, both points B and C in Figure 3 feature $\hat{S} = 1$, but none of them is an OSS (constant SE) because both points are operating at

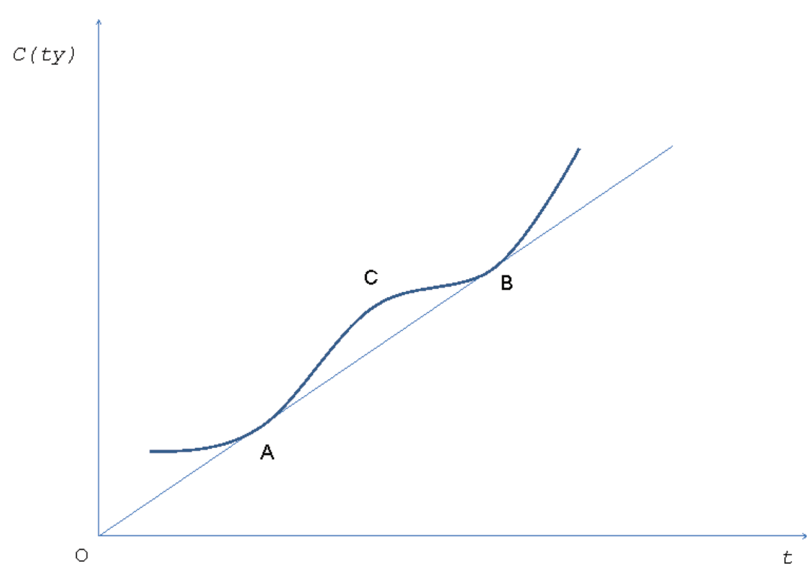

Figure 2. Nonconvex cost function with multiple RAC minima.

[10]From now on, we simplify notation by omitting w and y as independent variables in x.

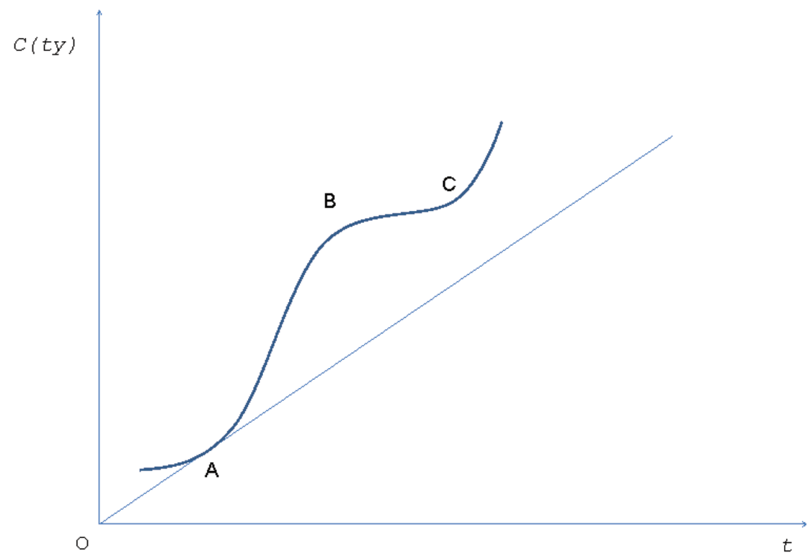

Figure 3. Nonconvex cost function with a unique RAC minimum.

decreasing SE. The same outcome occurs for point C in Figure 2: $\hat{S} = 1$, but it operates at sub-constant SE. In other words, in a nonconvex technology, local and global SE indicators may diverge.

RAC function in a convex technology

We now examine how convexity of the technology affects the shape of RAC. Lemma 1 is useful in ascertaining the behaviour of $C\left(t\boldsymbol{y}_j\right)$ as a function of t in the interval comprised between two OSS with $RAC^* > 0$. In fact, if the cost-efficient point j has two OSS, o and o' with finite $\gamma_{j,o}$ and $\gamma_{j,o'}$ respectively, this amounts to examining the behaviour of C in$\left[{}^{-}t, \bar{t}\right]$, where ${}^{-}t = \frac{1}{\gamma_{j,o}}$ and $\bar{t} = \frac{1}{\gamma_{j,o'}}$.

Theorem 1: *Under either assumptions (A.1–A.4) or (A.1bis)-(A.4), when $\boldsymbol{y}_j$ has two OSS with a finite γ, $RAC_{yj}{=}C\left(t\boldsymbol{y}_j\right)/t = RAC^* > 0 \quad t \in \left[{}^{-}t, \bar{t}\right]$.*

In graphical terms, Theorem 1 implies that for a convex technology the behaviour of the VRS cost function depicted in Figure 2 cannot occur, because in the interval between two OSS (i.e. points A and B), it does not deviate from the straight line OB denoting the CRS cost function, i.e. $t \cdot RAC^*$.

Remark 1: In case of a polyhedral technology Theorem 1 also holds for the interval $[0, \bar{t}]$, i.e. when $\gamma_{j,o} = \infty$ and $\gamma_{j,o'}$ is finite (see Appendix 2).

Theorem 1 establishes that in T the RAC function is constant in the interval between two OSS and it determines the conditions under which the 'flat-bottomed' RAC assumption holds (see Baumol, Panzar, and Willig 1982, 322–324). Therefore, given this result, we only need to determine the behaviour outside the above-mentioned range, that is to say the case where the output mix $\boldsymbol{y}_j$ is expanding/contracting towards the level determined by an OSS $(\boldsymbol{x}_o, \boldsymbol{y}_o), \frac{1}{\gamma_{j,o}} \cdot \boldsymbol{y}_j$ with $\gamma_{j,o} \neq 1$. To this purpose, we express any output level $\boldsymbol{y}' \in \left[\boldsymbol{y}_j, \frac{1}{\gamma_{j,o}} \cdot \boldsymbol{y}_j\right)$ as $\boldsymbol{y}' \equiv t\boldsymbol{y}_j \equiv (1-\alpha) \cdot \boldsymbol{y}_j + \alpha \cdot \left(\frac{1}{\gamma_{j,o}} \boldsymbol{y}_j\right)$, where $\alpha \in [0, 1)$ is a continuous control-parameter that determines t in the above range.[11] Using (8) and $\boldsymbol{y}' \equiv t\boldsymbol{y}_j$ – the RAC function can thus be expressed as:

$$RAC_{yj} = \frac{C\left(\boldsymbol{y}'\right)}{t} = \frac{\boldsymbol{w}_j \boldsymbol{x}_j^*(\alpha)}{1 + \alpha \cdot \left(\frac{1}{\gamma_{j,o}} - 1\right)} \tag{12}$$

We can now point out that the problem under discussion may be solved by analysing the derivative of expression (12) with respect to α, where a marginal increase in α denotes either an expansion ($\gamma_{j,o} < 1$) or a contraction ($\gamma_{j,o} > 1$) of output $\boldsymbol{y}'$

[11]The above interval holds for the case $\gamma_{j,o} < 1$, its bounds must be obviously reversed for the opposite case $\gamma_{j,o} > 1$.

towards the level determined by an OSS. This insight leads to the following key result.

Theorem 2: *Under either assumptions (A.1)-(A.4) or (A.1bis)-(A.4), RAC_{yj} is a monotonically decreasing/increasing function of t, respectively, at any $t < \frac{1}{\gamma_{j,o}}$ and $t > \frac{1}{\gamma_{j,o}}$. Moreover, this function is convex at any $t < \frac{1}{\gamma_{j,o}}$.*

We point out that Theorem 2 holds for both sets of assumptions: i.e. with or without $(0, 0)$ and also for the case $\gamma_{j,o} = 1$, i.e. when $\left(\boldsymbol{x}_j^*, \boldsymbol{y}_j\right)$ is an OSS. The algebraic proof only employs continuity of the cost function in outputs and Shephard's (1970) result on its convexity in outputs to determine the sign of the first and second right-derivatives of (12) with respect to α at $\alpha = 0$. These two properties ensure that the cost function is subdifferentiable in outputs, albeit we do not explicitly use subgradients in the proof, which holds for any convex technology including the case of nonparametric (polyhedral) production technologies.

The proof of Theorem 2 is based on $\delta(\alpha)$, the difference between $\boldsymbol{w}_j\boldsymbol{x}_j^*(\alpha)$ and its linear approximation $(1-\alpha)\cdot \boldsymbol{w}_j\boldsymbol{x}_j^* + \alpha \cdot \boldsymbol{w}_j\boldsymbol{x}_o$, where the cost function generated by a polyhedral technology corresponds to the case $\delta''(0) = 0$. Figure 4 illustrates the algebraic sign of $\delta'(0)$ and $\delta''(0)$ implied by the cost function of a convex technology. In Figure 4 (left-hand side), point A is a generic cost scale inefficient point, while C denotes its optimal scale size. The dashed line is the linear approximation of the cost function. Clearly, at point A: $\delta'(0) < 0$ and $\delta''(0) > 0$. Different from the smooth case, in Figure 4 (right-hand side) we have two kinds of cost scale inefficient points, i.e. point A and point B: for these, we, respectively, have $\delta'(0) < 0$ and $\delta'(0) = 0$.

Remark 2: Theorem 2 states that the RAC function is monotonically decreasing/increasing at any scale size lower/greater than the OSS, respectively.

We point out that Theorems 1 and 2 prove that in a convex technology, the RAC function is monotonic and convex at any output level which is strictly lower than its minimizer.

A full characterization of the RAC shape can be obtained only for the case of a polyhedral technology, because of the possibility of ascertaining the concavity of the RAC function for $t > \frac{1}{\gamma_{j,o}}$ (see proof of Theorem 2):

Corollary 1: Under assumptions (A.1 bis)-(A.4) and when the output mix $\boldsymbol{y}_j$ has no OSS which is a largest scale size, RAC_{yj} is a reversed-v function of t.

Note, in fact, that if an OSS is a largest scale size (i.e. the largest possible within the output bounds), then RAC does not feature an increasing branch. A noteworthy example satisfying the assumptions of Corollary 1 is the so-called BCC technology introduced by Banker, Charnes, and Cooper (1984) (see Section 6.1).

To the best of our knowledge, our results are new to the literature, as paradigmatically shown by Førsund and Hjalmarsson (2004) and Ray (2015) where a convex technology is assumed but no general conclusion on the properties of the RAC function is supplied. We here remark that the only result on the shape of RAC seems to be that of Färe

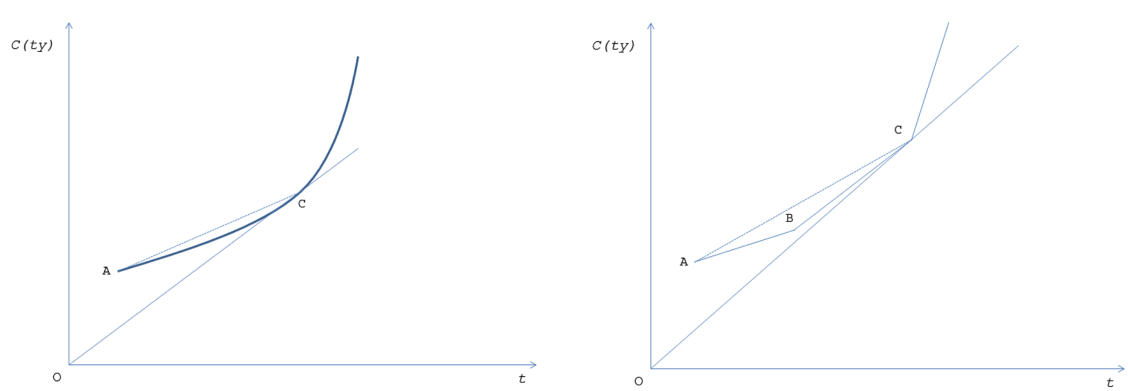

Figure 4. Cost function of a smooth technology (left hand side) and a polyhedral technology (right hand side).

(1974), regarding the single output case and where U-shape is meant to indicate monotonically decreasing/increasing RAC function (see Remark 2). In this respect, we point out that our contribution reveals that: a) the mentioned result carries over to the multi-output case[12]; b) strictly speaking, in a polyhedral technology, the concavity of a branch of the RAC function determines a reversed-v shape (as shown in Figure 5 in the case of multiple OSSs in $[t_1, t_2]$).

IV. Scale efficiency, primal and dual scale elasticities

In this section, the monotonicity of the RAC function is used to derive some fundamental properties concerning the relations among cost scale efficiency, primal and dual scale elasticities in a convex technology.

Cost scale efficiency and dual scale elasticity

The specific relationship between cost scale efficiency and dual scale elasticity is a result that is unavailable in the literature (see, e.g. Balk 2001; Färe and Primont 1995; Färe, Grosskopf, and Lovell 1986), as a consequence of the difficulties in the literature in determining the RAC shape (see Section 3.2).

The first step to get the needed result consists in the differentiation of RAC in expression (8) with respect to t evaluated at $t = 1$ yielding:

$$\left.\frac{dRAC(t)}{dt}\right|_{t=1} = C(\mathbf{y}) \cdot \left[\frac{1}{\hat{S}} - 1\right] \tag{13}$$

where $\hat{S}$ is the dual scale elasticity (5). At points where C is not differentiable, expression (13) takes two values determined by $\hat{S}^-$ and $\hat{S}^+$, respectively: the left and right limit at $\boldsymbol{y}$ of (5).

As a second step, we observe that the proof of Theorem 2 – part I – unambiguously determines the sign of (13) for cost scale inefficient points. Accordingly, we can formulate:

Proposition 2: Under assumptions (A.1bis)-(A.5), at a cost-efficient point $(\boldsymbol{x}^*, \boldsymbol{y}) \in T$ where C is differentiable, we have:

$$\begin{array}{lcl} \hat{S} > 1 & \Leftrightarrow & SCE < 1 \text{and GISE}, \\ \hat{S} = 1 & \Leftrightarrow & SCE = 1 \text{and GCSE}, \\ \hat{S} < 1 & \Leftrightarrow & SCE < 1 \text{and GDSE}. \end{array} \tag{14}$$

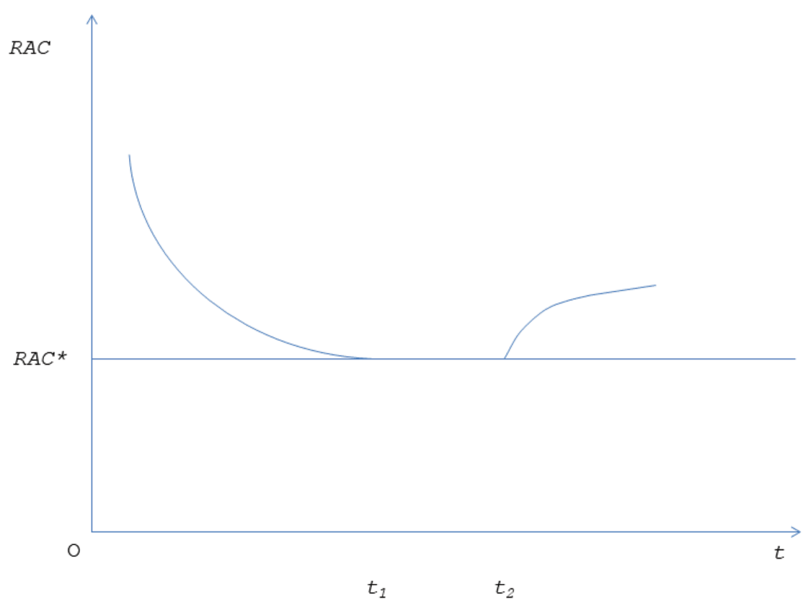

Figure 5. Reversed-v shape of RAC with multiple minima.

[12]Just like in Färe (1974), by excluding (0,0), Corollary 1 assumes that RTS can vary – along the efficient frontier – from increasing to decreasing when output increases.

where SCE is the scale efficiency measure (11) and GISE, GCSE and GDSE stand for globally increasing, globally constant, and globally decreasing scale economies, respectively.

At points where C is not differentiable but its left and right derivatives exist, the convexity of the cost function in the outputs implies that $\hat{S}$ is a non-increasing function,[13] of t. As a consequence, $\hat{S}^{-} \geq \hat{S}^{+}$: this result combined with Proposition 2 readily explains the following result:

Corollary 2: Under assumptions (A.1bis)-(A.5), at a cost-efficient point $(\boldsymbol{x}^{*}, \boldsymbol{y}) \in T$ where C is not differentiable, we have:

$$\begin{array}{rcl} \hat{S}^{-} \geq \hat{S} > 1 & \Leftrightarrow & SCE < 1 \text{and GISE}, \\ \hat{S}^{+} \leq 1 \leq \hat{S}^{-} & \Leftrightarrow & SCE = 1 \text{and GCSE}, \\ \hat{S}^{+} \leq \hat{S} < 1 & \Leftrightarrow & SCE < 1 \text{and GDSE}. \end{array} \tag{15}$$

Note that, with respect to Proposition 2, the GCSE case now includes three additional possibilities: (a) $\hat{S}^{+} < 1 < \hat{S}^{-}$, (b) $\hat{S}^{-} > \hat{S}^{+} = 1$, and (c) $\hat{S}^{+} < \hat{S}^{-} = 1$, i.e. C is not differentiable at a single OSS, and C is not differentiable at either the left or right end of the interval determined by multiple OSS points, respectively.

It can be easily proven that, under assumptions (A.1)-(A.5), the inclusion of $(0, 0)$ in T implies that only the last parts of Proposition 2 and Corollary 2 – regarding GDSE – hold for a smooth technology, while GCSE and GDSE occur for a polyhedral technology (see Remark 1 and Diewert 1973).

Proposition 2 and Corollary 2 conclude that in a convex technology both the scale-elasticity and the scale efficiency criterion provide the same indications for the SE regime. In other words, the result of Podinovski (2004) for local and global indicators of RTS for radial technical-efficient points on the production frontier is extended to the corresponding indicators of SE for efficient points along the cost function and to the non-differentiable case.[14] Moreover, the convexity of C and the RAC function enable to reach two additional significant properties (see Propositions 3 and 4 below) that cannot be obtained on the basis of the RAP function, which is only quasi-concave.

Primal and dual scale elasticities

Finally, we are in a position to generalize a result of Panzar and Willig (1977). In this regard, it suffices to consider Proposition 1 and the convexity of C in a convex technology to reach the following proposition:

Proposition 3: Under either assumptions (A.1)-(A.5) or (A.1bis)-(A.5), at a cost-efficient point $(\boldsymbol{x}^{*}, \boldsymbol{y}) \in T$ where C is not differentiable, we have

$$\begin{array}{c} S^{-} \geq \hat{S}^{-} \\ S^{+} \leq \hat{S}^{+} \end{array} \tag{16}$$

Proposition 3 reveals that $S \leq \hat{S}$ (i.e. proposition 5 in Panzar and Willig 1977, 491) only holds for a convex technology and for the right limits of primal and dual scale elasticities, while the sign of the inequality must be reversed for their left limits.

V. Correspondence between returns to scale and scale economies regimes

In a convex technology, the identification of scale economies regimes based on scale elasticity (i.e. local) and scale efficiency (i.e. global) measures, discussed in Section 4.1, permits to avoid reference to global measures and to focus directly on the relationship between primal and dual scale elasticities as indicators of returns to scale and scale economies regimes, respectively.

In a convex smooth technology, Proposition 1 confirms the well-known result about the coincidence of RTS and SE local measures. However, our Proposition 3 suggests that this simple result may not hold at points where the cost function is not differentiable. In this case, which indeed is expected to be the rule rather than the exception, as a consequence of both the axiomatic approach,

[13]According to definition (5), the dual scale elasticity at a point of the cost function corresponds to the ratio of the cosine of the angle between the line tangent to the point and the horizontal axis to the cosine of the angle of the ray joining the origin and the same point. From Figure 4, left and right, it can be easily seen that this scale elasticity is a decreasing function of t in the smooth case (left) and a non-increasing function of t in the polyhedral technology (right).

[14]See Theorem 7 in Podinovski (2004) where non-differentiability is not considered. Moreover, his conclusions cannot be applied to the cost function except in the tautological case of a homothetic technology (see, ibid., p 241).

which does not ensure differentiability to hold everywhere, and of the greater generality of the discontinuous differentiability assumption[15], a specific relationship must be worked out. To this purpose, Corollary 2 and Proposition 3 can be employed to yield the following result:

Proposition 4: Under assumptions (A.1bis)-(A.5), at a cost-efficient point $(\boldsymbol{x}^*, \boldsymbol{y}) \in T$ we have:

$$\begin{array}{ccc} \text{ISE} & \Rightarrow & \text{IRS or CRS} \\ \text{CSE} & \Rightarrow & \text{CRS} \\ \text{DSE} & \Rightarrow & \text{DRS or CRS} \end{array} \tag{17}$$

where IRS, CRS and DRS stand for increasing, constant and decreasing returns to scale, respectively.

We remark that Proposition 4 can also be read from the right to the left side: in fact, it can be immediately checked that, at a cost-efficient point $(\boldsymbol{x}^*, \boldsymbol{y}) \in T$ we have:

$$\begin{array}{ccc} \text{IRS} & \Rightarrow & \text{ISE} \\ \text{CRS} & \Rightarrow & \text{ISE or CSE or DSE} \\ \text{DRS} & \Rightarrow & \text{DSE} \end{array} \tag{18}$$

Remark 3: Under assumptions (A.1)-(A.5), the inclusion of $(0, 0)$ in T implies that only the last part of Proposition 4 – regarding DSE – holds for a smooth technology, while only DSE and CSE are possible for a polyhedral technology.

Relationships (17) and (18) are important since, to the best of our knowledge, such kind of correspondence for general convex technologies is not found in the literature. Therefore, we remark that non-differentiability of the cost function can bring about a significant divergence between RTS and SE regimes even in a convex setting, where local and global indicators are coincident. This result complements at the local level some recent findings proving the differences at the global level, i.e. when using measures based on scale efficiency, between RTS and SE concepts (strict scale economies and scale economies in the jargon of Baumol 1977) in non-homothetic production technologies, as those considered in Cesaroni and Giovannola (2015) and Zelenyuk (2014).

Our findings are important for a variety of issues concerning resource allocation, competition policy and regulation within the industrial and service sectors, which are discussed in Appendix 5. Moreover, this same appendix deals with the application of our results to duality studies within the parametric approach to frontier estimation.

VI. Empirical illustration

As argued in Appendix 5.2, a convex nonparametric estimation method ensures a greater ease and precision in testing the empirical validity of our theoretical results. More specifically, we employ two DEA models: 1) the convex reference technology of the BCC model, which satisfies assumptions (A.1bis)-(A.5), 2) its modified version satisfying assumptions (A.1)-(A.5).

We consider n observations indexed by $j(j = 1, \ldots, n)$, the observed input and output vectors are $\left(\boldsymbol{x}_j, \boldsymbol{y}_j\right)$ while the $m \times n$ matrix of inputs and the $s \times n$ matrix of outputs are, respectively, denoted as $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_n]$ and $\mathbf{Y} = [\mathbf{y}_1, \ldots, \mathbf{y}_n]$. In the BCC model, model 1, we have $\boldsymbol{x}_j, \boldsymbol{y}_j \neq 0$ and the corresponding technology can be described as:

$$T = \left\{ (\boldsymbol{x}, \boldsymbol{y}) \middle| \mathbf{X}\lambda \leq \mathbf{x}, \mathbf{Y}\lambda \geq \mathbf{y}, \lambda_j \geq 0, \sum_{j=1}^{n} \lambda_j = 1 \right\}, \tag{19}$$

where λ is the $n \times 1$ vector with components equal to λ_j. Clearly, a point $(\boldsymbol{x}, \boldsymbol{y})$ for which inequalities in (19) are strict is an inefficient point. Remark that this technology is a VRS technology since it allows for IRS, DRS and CRS to be present along the efficient frontier, thus permitting a complete analysis of the scale efficiency. Moreover, it is the least restrictive technology since it generates the smallest set enveloping all original observations.

Conversely, model 2 is obtained from model 1 by including the fictitious observation **(0,0)** in technology (19). This changes the nature of the efficient frontier, which now exhibits non-increasing returns-to-scale (NIRS: see Remark 3). The following subsections extensively illustrate the

[15]The relevance of non-differentiability of the cost function in the axiomatic approach is empirically shown in Section VI Panzar and Willig (1977, 488) admit that continuous differentiability of the cost function is a strong assumption.

empirical results from the more general technology of model 1, while those relating to model 2 are shown in Appendix 6.

Model 1

The estimates refer to the dataset of Cesaroni, Kerstens, and Van de Woestyne (2017). These data are published as supplementary material and are freely accessible at: https://link.springer.com/article/10.1057/s41274-016-0162-7. It concerns a three-input-two-output technology associated with the Italian local-public-transit sector, where each firm faces a specific input price vector.

Corollary 2 of Proposition 2 is shown in Table 2 where our RAC method for the estimation of global SE is coupled with that of Sueyoshi (1999) for cost scale elasticities $\hat{S}^-$ and $\hat{S}^+$, both methods being applied to cost-efficient points, i.e. to the projections of the original observations on to the cost frontier. For the sake of completeness, OE denotes the VRS cost efficiency of the original observation. In particular, we spell out the details of all elements reported in Table 2. OE is obtained by substituting the numerator of expression (11) with the optimal cost in the VRS technology. The Rj ratio is given by expression (11), while the gamma ratio $\gamma_{j,o}$ is defined by expression (10) and indicates increasing/constant/decreasing global SE when lower/equal/greater than one, respectively. These variables have been computed by means of the R-package Benchmarking. Cost scale elasticities $\hat{S}^-$ and $\hat{S}^+$ are computed by means of models (34) and (30) in Sueyoshi (1999). These computations have been performed in Maple.

We remark that computations empirically confirm the coincidence of global and local scale economies indicator (GSE and LSE) in the case of a convex technology at points where C is not differentiable, as established by Corollary 2, and also at points where C is differentiable, as established by Corollary 1. As for decision support, note that $\gamma_{j,o}$ offers useful information which complements the one obtained from the dual scale elasticity (ratio of average to marginal cost), permitting to compare discrete and marginal variations in production. For example, DMU 36 features an average elasticity of 1.224 (= (1.267 + 1.181)/2) at the current level of production, and a discrete variation in production of nearly 100% $\left(1/\gamma - 1\right)$ with respect to its OSS. Thus, while a marginal increase in production is clearly profitable – due to negligible adjustment costs –, by contrast these could be quite relevant if the DMU decided to implement the discrete variation in production. Note, e.g. that such a dilemma does not affect DMU 21. Finally, we note that the cost function is not differentiable at 36 points out of 43, as clearly indicated by the diverging values of the left ($\hat{S}^-$) and right ($\hat{S}^+$) cost scale elasticities in Table 2.

Table 2. Global and local scale economies (RAC and Sueyoshi methods) Model 1.

	Global scale economies				Local scale economies		
DMU	OE	R_j^*	$\gamma_{j,o}$	GSE	$\hat{S}^-$	$\hat{S}^+$	LSE
1	0.636	0.998	0.976	I#	1.141	1.095	I
2	1.000	0.931	4.546	D#	0.981	0.000	D
3	0.564	0.992	1.157	D	0.823	0.823	D
4	0.579	0.929	0.553	I	1.286	1.190	I
5	0.718	0.794	4.616	D	0.225	0.225	D
6	0.713	0.975	0.799	I	1.201	1.143	I
7	0.629	0.994	0.940	I	1.163	1.111	I
8	0.627	0.975	1.535	D	0.955	0.848	D
9	1.000	1.000	1.000	C#	1.186	0.701	C
10	0.605	0.858	0.366	I	1.486	1.288	I
11	0.795	0.946	2.976	D	0.973	0.910	D
12	0.664	0.861	0.347	I	1.508	1.269	I
13	0.467	0.967	1.733	D	0.957	0.854	D
14	0.635	0.915	0.480	I	1.322	1.195	I
15	0.533	0.945	3.994	D	0.982	0.934	D
16	0.660	0.991	0.915	I	1.176	1.122	I
17	0.971	0.915	0.523	I	1.329	1.218	I
18	1.000	0.809	0.304	I	+inf	1.379	I
19	0.738	0.997	0.971	I	1.146	1.098	I
20	0.699	0.981	0.820	I	1.172	1.118	I
21	0.737	0.999	1.018	D	0.932	0.932	D
22	1.000	0.512	16.000	D	0.906	0.530	D
23	0.928	0.934	4.230	D	0.980	0.931	D
24	0.380	0.940	2.868	D	0.969	0.904	D
25	0.664	0.868	0.393	I	1.428	1.278	I
26	0.530	0.989	0.868	I	1.146	1.094	I
27	0.429	0.959	2.251	D	0.968	0.892	D
28	0.756	0.972	1.524	D	0.949	0.830	D
29	0.878	0.874	4.568	D	0.600	0.200	D
30	0.562	0.949	0.614	I	1.232	1.151	I
31	0.717	0.991	1.150	D	0.816	0.815	D
32	1.000	1.000	1.000	C	1.149	0.781	C
33	0.545	0.936	0.549	I	1.257	1.164	I
34	1.000	0.555	17.564	D	0.922	0.000	D
35	0.795	0.956	1.866	D	0.952	0.829	D
36	0.577	0.925	0.509	I	1.267	1.181	I
37	0.736	0.989	1.165	D	0.936	0.936	D
38	0.649	0.982	0.820	I	1.162	1.109	I
39	0.875	0.601	8.174	D	0.434	0.434	D
40	0.735	0.936	3.101	D	0.971	0.910	D
41	0.641	0.979	1.617	D	0.967	0.862	D
42	0.543	0.953	4.288	D	0.942	0.942	D
43	0.818	0.952	2.452	D	0.968	0.892	D

#I = Increasing scale economies; C = Constant scale economies; D = Decreasing scale economies.

Table 3 employs the results obtained from the application of Sueyoshi (1999) computation method regarding primal (S^-,S^+) and dual ($\hat{S}^-$,$\hat{S}^+$) scale elasticities, respectively, to show the empirical effectiveness of Propositions 3 and 4 in establishing the correspondence between RTS and SE classifications. In particular, the primal scale elasticities are obtained from models (33) and (14) in Sueyoshi (1999), while the dual scale elasticity is computed as indicated when discussing Table 2. Computations are performed in Maple.

Table 3 makes it evident that inequalities (16) of Proposition 3 empirically hold, but also that the correspondences (17) and (18) (associated to Proposition 4) are satisfied (see columns RTS and SE). Moreover, the coincidence between local and global RTS, proven explicitly in Podinovski (2017), and that between local and global SE (proven by our Proposition 3) allow to conclude that global RTS and global SE differ for 37 out of 43 cost-efficient points examined in Table 3. This proves the absolute relevance of one of the main motivations of this research and emphasizes the specific importance of RAC in general, i.e. for non-homothetic production technologies.

We point out that the empirical validity of Proposition 4 confirms that Sueyoshi's (1999) interpretation of the relationship between RTS and SE classification is not appropriate. Moreover, it is interesting to evaluate differences in RTS versus SE classification brought about by the primal and the dual scale elasticity in the presence of non-differentiability. In this regard, two remarks are in order. First, as far as qualitative information is concerned, 37 out of 43 DMUs exhibit a conflicting classification (compare columns 4 and 7 in Table 3). Second, from the quantitative point of view, the average difference between left and right scale elasticities is equal to 0.437 for the primal, and 0.146 for the dual. Thus, the cost criterion yields narrower estimates of the scale elasticity when compared to the production criterion. This testifies that the breakdown of the standard duality correspondence may be quite significant from the point of view of the policy prescriptions ensuing from RTS and SE analyses, respectively.

Finally, Figure 6 illustrates the different shapes that VRS and CRS cost and RAC functions assume in convex (C) and nonconvex (NC) nonparametric technologies. This is being done with reference to the output mix of observation 21: Figure 6 (left-hand side) plots $C(ty)$ against t (just like in Figures 2 and 3), while Figure 6 (right-hand side) plots RAC against t. Starting from positive values of t near 0 and expanding the output mix towards 1, the convex VRS (C-VRS) technology exhibits a reversed-ν RAC. By contrast, in addition to the discontinuities generated by the horizontal segments of its cost function, the nonconvex VRS (NC-VRS) technology presents a RAC which is not of the reverse-ν kind, because it is simultaneously increasing/decreasing both for $t < 1$ and $t > 1$.

Table 3. Primal and dual scale elasticities, RTS and SE. Model 1.

	Primal scale el. and RTS			Dual scale el. and SE		
DMU	S^-	S^+	RTS	$\hat{S}^-$	$\hat{S}^+$	SE
1	1.270	0.993	C	1.141	1.095	I
2	1.005	0.000	C	0.981	0.000	D
3	1.017	0.732	C	0.823	0.823	D
4	1.539	0.986	C	1.286	1.190	I
5	0.944	0.102	D	0.225	0.225	D
6	1.341	0.991	C	1.201	1.143	I
7	1.282	0.993	C	1.163	1.111	I
8	1.013	0.783	C	0.955	0.848	D
9	1.300	0.644	C	1.186	0.701	C
10	1.965	0.976	C	1.486	1.288	I
11	1.007	0.874	C	0.973	0.910	D
12	2.049	0.974	C	1.508	1.269	I
13	1.012	0.802	C	0.957	0.854	D
14	1.652	0.984	C	1.322	1.195	I
15	1.005	0.903	C	0.982	0.934	D
16	1.291	0.993	C	1.176	1.122	I
17	1.580	0.985	C	1.329	1.218	I
18	+inf	0.968	C	+inf	1.379	I
19	1.271	0.993	C	1.146	1.098	I
20	1.331	0.992	C	1.172	1.118	I
21	1.019	0.890	C	0.932	0.932	D
22	0.985	0.336	D	0.906	0.530	D
23	1.005	0.908	C	0.980	0.931	D
24	1.007	0.870	C	0.969	0.904	D
25	1.868	0.979	C	1.428	1.278	I
26	1.309	0.992	C	1.146	1.094	I
27	1.009	0.840	C	0.968	0.892	D
28	1.013	0.782	C	0.949	0.830	D
29	0.936	0.082	D	0.600	0.200	D
30	1.471	0.988	C	1.232	1.151	I
31	1.017	0.731	C	0.816	0.815	D
32	1.262	0.703	C	1.149	0.781	C
33	1.545	0.986	C	1.257	1.164	I
34	1.006	0.000	C	0.922	0.000	D
35	1.012	0.800	C	0.952	0.829	D
36	1.602	0.985	C	1.267	1.181	I
37	1.017	0.733	C	0.936	0.936	D
38	1.331	0.992	C	1.162	1.109	I
39	0.971	0.233	D	0.434	0.434	D
40	1.007	0.878	C	0.971	0.910	D
41	1.013	0.791	C	0.967	0.862	D
42	1.005	0.909	C	0.942	0.942	D
43	1.009	0.851	C	0.968	0.892	D

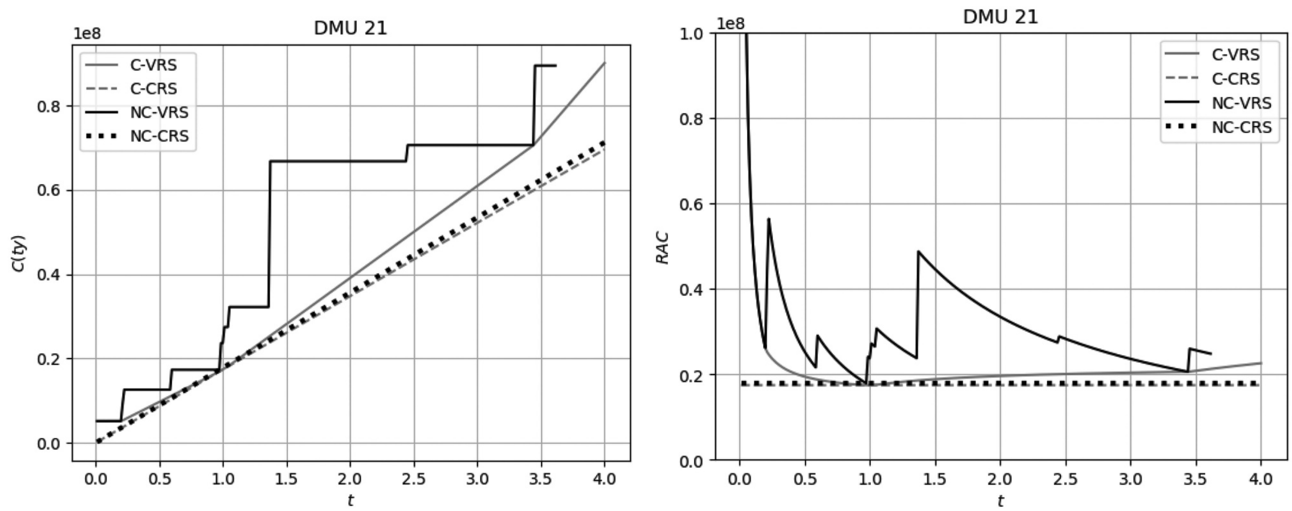

Figure 6. Cost function C(ty) (left) and RAC (right) for DMU 21.

Model 2

The empirical results relating to model 2 are available in Appendix 6.

VII. Conclusions

This contribution advances the literature on scale economies in convex technologies in several ways. First, it determines the shape of the RAC function and extends the coincidence of local and global returns to scale from production to cost analysis and to the case of multiple minima and with a non-differentiable cost function. Second, it introduces a global method for the determination of scale economies which simplifies the Färe and Grosskopf (1985) procedure and supplies qualitative and quantitative information, by means of the $\gamma_{j,o}$ coefficient, that is otherwise not available in the axiomatic approach. Third, besides showing that a required monotonicity condition of the local degree of homogeneity of the technology is satisfied on the basis of the former results, it establishes a correct correspondence between returns to scale and scale economies classification for points of the cost function where differentiability does not hold. An empirical illustration documents these results.

In nonconvex technologies, the complexity of the analysis – as revealed by Proposition 1bis – calls for a more specific inquiry which is left for future research.

Disclosure statement

No potential conflict of interest was reported by the author(s).

References

Balk, B. M. 2001. "Scale Efficiency and Productivity Change." *Journal of Productivity Analysis* 15 (3): 159–183. https://doi.org/10.1023/A:1011117324278.

Banker, R. D. 1984. "Estimating Most Productive Scale Size Using Data Envelopment Analysis." European Journal of Operational Research 17 (1): 35–44. https://doi.org/10.1016/0377-2217(84)90006-7.

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 (3): 473–481. https://doi.org/10.1016/0377-2217(95)00044-5.

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 (9): 1078–1092. https://doi.org/10.1287/mnsc.30.9.1078.

Banker, R. D., and R. M. Thrall. 1992. "Estimation of Returns to Scale Using Data Envelopment Analysis." European Journal of Operational Research 62 (1): 74–84. https://doi.org/10.1016/0377-2217(92)90178-C.

Baumol, W. J. 1977. "On the Proper Cost Tests for Natural Monopoly in a Multiproduct Industry." American Economic Review 67:809–822.

Baumol, W. J., and D. Fischer. 1978. "Cost-Minimizing Number of Firms and Determination of Industry Structure." Quarterly Journal of Economics 92 (3): 439–468. https://doi.org/10.2307/1883153.

Baumol, W. J., J. C. Panzar, and R. D. Willig. 1982. *Contestable Markets and the Theory of Industry Structure*. New York: Harcourt Brace Jovanovich.

Berger, A. N. 1993. “Distribution Free Estimates of Efficiency in the U.S. Banking Industry and Tests of the Standard Distributional Assumptions.” Journal of Productivity Analysis 4 (3): 261–292. https://doi.org/10.1007/BF01073413.

Cassels, J. M. 1937. “Excess Capacity and Monopolistic Competition.” Quarterly Journal of Economics 51 (3): 426–443. https://doi.org/10.2307/1884835.

Cesaroni, G., and D. Giovannola. 2015. “Average-Cost Efficiency and Optimal Scale Sizes in Non-Parametric Analysis.” European Journal of Operational Research 242 (1): 121–133. https://doi.org/10.1016/j.ejor.2014.09.062.

Cesaroni, G., K. Kerstens, and I. Van de Woestyne. 2017. “Estimating Scale Economies in Non-Convex Production Models.” Journal of the Operational Research Society 68 (11): 1442–1451. https://doi.org/10.1057/s41274-016-0162-7.

Chambers, R. G. 1988. *Applied Production Analysis: A Dual Approach*. Cambridge: Cambridge University Press.

Chavas, J. P., and T. L. Cox. 1999. “A Generalized Distance Function and the Analysis of Production Efficiency.” Southern Economic Journal 66 (2): 294–318. https://doi.org/10.1002/j.2325-8012.1999.tb00248.x.

Coelli, T., and S. Perlman. 1999. “A Comparison of Parametric and Non-Parametric Distance Functions: With Application to European Railways.” European Journal of Operational Research 117 (2): 326–339. https://doi.org/10.1016/S0377-2217(98)00271-9.

Diewert, W. E. 1973. “Functional Forms for Profit and Transformation Functions.” Journal of Economic Theory 6 (3): 284–316. https://doi.org/10.1016/0022-0531(73)90051-3.

Färe, R. 1974. “A Necessary and Sufficient Condition for a U-Shaped Average Cost Function.” The Swedish Journal of Economics 76 (2): 236–240. https://doi.org/10.2307/3439205.

Färe, R., and S. Grosskopf. 1985. “A Nonparametric Cost Approach to Scale Efficiency.” Scandinavian Journal of Economics 87 (4): 594–604. https://doi.org/10.2307/3439974.

Färe, R., S. Grosskopf, and C. A. K. Lovell. 1983. “The Structure of Technical Efficiency.” Scandinavian Journal of Economics 85 (2): 181–190. https://doi.org/10.2307/3439477.

Färe, R., S. Grosskopf, and C. A. K. Lovell. 1986. “Scale Economies and Duality.” Journal of Economics 46 (2): 175–182. https://doi.org/10.1007/BF01229228.

Färe, R., S. Grosskopf, and C. A. K. Lovell. 1988. “Scale Elasticity and Scale Efficiency.” Journal of Institutional and Theoretical Economics 144:721–729.

Färe, R., and D. Primont. 1995. *Multi-Output Production and Duality: Theory and Applications*. Norwell: Kluwer.

Ferrier, G. D., and C. A. K. Lovell. 1990. “Measuring Cost Efficiency in Banking: Econometric and Linear Programming Evidence.” Journal of Econometrics 46 (1–2): 229–245. https://doi.org/10.1016/0304-4076(90)90057-Z.

Førsund, F., and L. Hjalmarsson. 2004. “Are All Scales Optimal in DEA? Theory and Empirical Evidence.” Journal of Productivity Analysis 21 (1): 25–48. https://doi.org/10.1023/B:PROD.0000012451.84910.6e.

Frisch, R. 1965. *Theory of Production*. Dordrecht: Reidel Publishing Company.

Koopmans, T. C. 1957. *Three Essays on the Nature of Economic Science*. New York: McGraw-Hill.

McFadden, D. 1978. “Cost, Revenue and Profit Functions.” In *Production Economics: A Dual Approach to Theory and Applications*, edited by M. A. Fuss and D. McFadden. Amsterdam: North-Holland 3-–109.

Panzar, J. C., and R. D. Willig. 1977. “Economies of scale in multi-output production.” Quarterly Journal of Economics 91 (3): 481–493. https://doi.org/10.2307/1885979.

Podinovski, V. 2004. “Efficiency and Global Scale Characteristics on the ?no Free Lunch? Assumption Only.” Journal of Productivity Analysis 22 (3): 227–257. https://doi.org/10.1007/s11123-004-7575-z.

Podinovski, V. 2017. “Returns to Scale in Convex Production Technologies.” European Journal of Operational Research 258 (3): 970–982. https://doi.org/10.1016/j.ejor.2016.09.029.

Ray, S. 2015. “Nonparametric Measures of Scale Economies and Capacity Utilization: An Application to U.S. Manufacturing.” European Journal of Operational Research 245 (2): 602–611. https://doi.org/10.1016/j.ejor.2015.03.024.

Shephard, R. W. 1970. *Theory of Cost and Production Functions*. Princeton: Princeton University Press.

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. https://doi.org/10.1287/mnsc.45.11.1593.

Tone, K., and B. K. Sahoo. 2003. “Scale, Indivisibilities and Production Function in Data Envelopment Analysis.” International Journal of Production Economics 84 (2): 165–192. https://doi.org/10.1016/S0925-5273(02)00412-7.

Zelenyuk, V. 2014. “Scale Efficiency and Homotheticity: Equivalence of Primal and Dual Measures.” Journal of Productivity Analysis 42 (1): 15–24. https://doi.org/10.1007/s11123-013-0361-z.