IÉSEG SCHOOL OF MANAGEMENT

May 2020

WORKING PAPER SERIES

2020-EQM-03

The Shape of Ray Average Cost and Its Role in Multioutput Scale Economies: Some Generalisations

Giovanni Cesaroni
Department for public administration, Prime Minister's Office, Rome, Italy

Kristiaan Kerstens
IÉSEG School of Management and LEM-CNRS (UMR 9221), Lille, France

Ignace Van de Woestyne
KU Leuven, Research Unit MEES, Brussels, Belgium

IÉSEG School of Management Lille Catholic University 3, rue de la Digue
F-59000 Lille
Tel: 33(0)3 20 54 58 92 www.ieseg.fr

The Shape of Ray Average Cost and Its Role in Multioutput Scale Economies: Some Generalisations

Giovanni Cesaroni[†]

Kristiaan Kerstens[‡]

Ignace Van de Woestyne[§]

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 reformulate Panzar and Willig's (1977) result 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. These general results are then applied to simplification and solution of some theoretical and computational problems featured by some important models for the estimation of scale economies, such those of Baumol and Fisher (1978), Färe and Grosskopf (1985) and Sueyoshi (1999).

Keywords: Returns to scale, Scale economies, Ray average cost

JEL Codes: C61, D24, L25

12 May 2020

† Department for public administration, Prime Minister's Office, Via del Sudario 49, I-00186 Rome, Italy. g.cesaroni@governo.it

‡ IÉSEG School of Management, CNRS-LEM (UMR 9221), Univ. Lille, 3 rue de la Digue, FR-59000 Lille, France. k.kerstens@ieseg.fr Corresponding author.

§ KU Leuven, Research unit MEES, Warmoesberg 26, BE-1000 Brussels, Belgium. ignace.vandewoestyne@kuleuven.be

1. Introduction

The introduction of the concept of ray average cost (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 et al. (1982). Following Färe et al. (1988, pp. 721-722), these contributions fall within the "neoclassical approach" to the measurement of scale economies, 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 and dual scale elasticities. Conversely, 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 et al. 1988, p. 721), with the comparison yielding qualitative information on the nature of scale economies (see, e.g., Färe et al. 1983, Banker et al. 1984). This approach is based on the computation of scale efficiency measures and generally disregards the local behavior 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 concerned with global measures[1] in returns to scale analysis in production: i.e., on the detection of the direction towards the most productive scale size (see Banker 1984).

A first attempt to establish a correspondence between scale elasticity and scale efficiency measures is offered in Färe et al. (1988). However, their results are limited to the analysis of returns to scale - characterized by equiproportionate changes in inputs and outputs - in the specific case in which the regular ultra-passum law holds (see Frisch (1965)). Moreover, these authors fail to realize that the definition of the degree of scale economies taken from Panzar and Willig (1977) is not valid for any behavior of the coefficient *r* along the production frontier, a problem that the literature has

[1] The adjective global is introduced in Podinovski (2004).

not yet addressed. A second important contribution for the integration of both approaches is made by Chavas and Cox (1999, p. 308). They are the first to explicitly define the RAC function in the context of production possibilities sets 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, p. 812). However, the impossibility of establishing the behavior of the RAC function in a general technology, an unsolved issue in the neoclassical approach (see, e.g., Baumol 1977, Baumol and Fisher 1978), led these authors to conclude that their proposed cost scale efficiency measure is a criterion to detect the presence of local CRS.

The recourse to cost analysis is required by the necessity of taking into account of the change in input proportions associated to scale effects, an element absent in the more traditional analysis of the axiomatic approach based on returns to scale (RTS), where only the ratio between the proportional change in outputs to that in the inputs - i.e., the ray average productivity (RAP) - is considered (see Banker and Thrall 1992, Banker et al. 1996, section 2.1 in Tone and Sahoo 2003). In fact, as shown in Cesaroni and Giovannola (2015) and Zelenyuk (2014), the presence of allocative inefficiency in input proportions determines different global classification results between RAC and RAP analyses. This clarifies why Podinovski's (2004) result on the monotonic behavior of the RAP function in a convex technology may not apply to RAC, and as a consequence the coincidence between local and global RTS indicators may not necessarily hold for scale economies (SE) indicators.

For these reasons, our work pursues a better integration of the two 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: the determination of the shape of the RAC, the relation between cost scale efficiency and dual scale elasticity measures, and the correspondence

between primal and dual scale elasticity measures when differentiability of the cost function does not hold.

This contribution is structured as follows. Section 2 introduces assumptions and some preliminary definitions, along with a critical discussion of the duality relationships between primal and dual scale elasticities of Panzar and Willig (1977). Section 3 derives the shape of RAC in the presence of multiple optima on the basis of a Shepard's (1970) property of the cost function in the output space. Section 4 shows the ensuing relationships between cost scale efficiency, and primal and dual scale elasticities. Section 5 applies some results of the previous section to determine the exact relation between RTS and SE classifications in a convex technology. Section 6 applies our theoretical results to the solution of problems featured in some important models for the determination of SE, such those of Färe and Grosskopf (1985), Sueyoshi's (1999) and Baumol and Fisher (1978). In particular, it is shown that Sueyoshi's characterization of the relation between RTS and SE is incorrect. Section 7 illustrates our main results by means of an empirical application based on a secondary data set using a non-parametric multiple inputs and outputs technology. Finally, Section 8 summarizes our contributions to the literature.

2. Production Technology and Duality between Scale Elasticity Measures

2.1 Definitions and Review

The multiple input and output production process is described as an $m \times 1$ input vector $\mathbf{x} \in \mathbf{X}$ used in the production of an $s \times 1$ output vector $\mathbf{y} \in \mathbf{Y}$, where $\mathbf{X} \subset \mathbf{R}_{+}^{\mathbf{m}}$ and $\mathbf{Y} \subset \mathbf{R}_{+}^{s}$ are compact sets. The production technology, or production possibility set, is represented by the feasible set

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

where T is a closed subset of $\mathbf{X} \times \mathbf{Y}$ which we assume to satisfy some standard regularity conditions (see, e.g., Panzar and Willig 1977, Färe and Primont 1995, Chavas and Cox 1999):

(A.1) T is non-empty and there exists $\mathbf{x} \geq \mathbf{0}$ and $\mathbf{y} \geq \mathbf{0}$ such that $(\mathbf{x}, \mathbf{y}) \in T$;

(A.2) $(\mathbf{0}, \mathbf{y}) \notin T$, unless $\mathbf{y} = \mathbf{0}$;

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

The above conditions state that in the feasible set, which includes its boundary and is bounded, (i) there are some semi-positive production processes, (ii) no free lunch is allowed and inaction is possible, (iii) inputs and outputs are freely (strongly) disposable. These three assumptions are maintained throughout the contribution, with T representing a general, nonconvex, feasible set. When convexity is required, then the feasible set is denoted by T_C .

According to Panzar and Willig (1977, pp. 487), conditions (A.1)-(A.3) are sufficient to establish that efficient production exists and that a continuous production transformation function

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

with $\frac{\partial \Phi}{\partial x_i} \geq 0$ and $\frac{\partial \Phi}{\partial y_j} \leq 0$ exists - where i and j denote a generic input and output, respectively. The 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_i and outputs y_j , where

$$C(\mathbf{y}, \mathbf{w}) = min\{\mathbf{w}\mathbf{x} \mid (\mathbf{x}, \mathbf{y}) \in T\}, \tag{3}$$

with $\mathbf{w} \in \mathbf{R}^m_{++}$ a vector of input prices.

In the remainder of the contribution, an additional assumption is sometimes considered:

(A.4) $\Phi(\mathbf{x},\mathbf{y})$ is continuously differentiable[2] in $\mathbf{x}$, and in y_j for $y_j > 0$ at points $(\mathbf{x},\mathbf{y})$ where $\mathbf{x}$ is cost-efficient for $\mathbf{y}$;

As far as RTS are concerned, observe that no assumption is being imposed on both *T* and T_C, which therefore are variable returns to scale (VRS) technologies. For future reference, we remark that the CRS extension of technology *T* is $T^{CRS} = \{(\mathbf{x},\mathbf{y}) : (\lambda\mathbf{x},\lambda\mathbf{y}) \in T \text{ for some } \lambda > 0\}$.[3] Finally, note that the smoothness condition (A.4) is sufficient for the existence of the partial derivatives of *C*: $C_j \equiv \dfrac{\partial C(.)}{\partial y_j}$ at any $y_j > 0$ (see Panzar and Willig 1977, p. 488).

In the framework built by assumptions (A.1)-(A.4), Panzar and Willig (1977) are the first to prove two important results on the relationship between production and cost scale elasticities in multi-output production, where the first can be defined as the ratio of the maximum proportional expansion in outputs associated to a given proportional expansion in inputs, and the second as the ratio of average to marginal cost (see also Färe and Primont 1995, pp. 39 and 53, and Tone and Sahoo 2003, p. 173). Panzar and Willig (1977) formulate these scale elasticities in a general and useful way that applies not only to the standard case of a smooth technology. More specifically, primal and dual scale elasticities are respectively defined as:

Definition 1: Production scale elasticity is defined as:

$$S = sup\{r \mid \exists \delta > 1 \text{ such that } (\lambda\mathbf{x}, \lambda^r\mathbf{y}) \in T \text{ for } 1 \le \lambda \le \delta\}. \qquad (4)$$

Definition 2: Cost scale elasticity is defined as:

$$\hat{S} = sup\{r \mid \exists \delta > 1 \text{ such that } C(\lambda\mathbf{y},\mathbf{w}) \le \lambda^{1/r} C(\mathbf{y},\mathbf{w}) \text{ for } 1 \le \lambda \le \delta\}. \qquad (5)$$

[2] Assumption R3 in Panzar and Willig (1977, pp. 487-488).

[3] Obviously, the same holds - *mutatis mutandum* - for the CRS extension of the convex technology.

On this basis, at a cost-efficient point $\mathbf{x}^*(\mathbf{w},\mathbf{y})$, where $C(\mathbf{y},\mathbf{w}) = \mathbf{w}\mathbf{x}^*$, Panzar and Willig (1977) prove the following:

Result 1): Under (A.4), equality $S = \hat{S}$ obtains;

Result 2): Without (A.4), the inequality relationship $S \leq \hat{S}$ holds.

Here, we remark that these pioneering results are only correct under specific conditions. First, definitions (4) and (5) may not necessarily apply to either *T* or T_C. Second, Result 1) is not properly formulated, not only for the previous reason, but also because it may be shown to imply $S \geq \hat{S}$ (see Appendix 1). Third, since definitions (4) and (5) exclude values $\lambda \leq 1$, which is legitimate only in the smooth case, Result 2) is an incomplete description of the relationships between S and $\hat{S}$ at a frontier point of the feasible set T_C where continuous differentiability of $\Phi(.)$ and differentiability of *C* are not possible.

For the purposes of the present contribution, a first step is achieved by means of the refinement of Result 1, which we are developing in the following sub-section.

2.2 A Duality Relationship

As mentioned earlier, for Result 1 to be referable to a general feasible set *T*, one needs to employ a different operational definition of primal and dual scale elasticities compared to (4) and (5). Given the use of assumption (A.4), these elasticities can be defined in a straightforward manner by employing differential calculus as

Definition 3:

Production scale elasticity under (A.4) is

$$S \equiv -\frac{\sum_i x_i \, \partial\Phi / \partial x_i}{\sum_j y_j \, \partial\Phi / \partial y_j}. \qquad (4bis)$$

Definition 4:

Cost scale elasticity under (A.4) is

$$\hat{S} \equiv \frac{C(\mathbf{y}, \mathbf{w})}{\sum_j y_j C_j(\mathbf{y}, \mathbf{w})}. \qquad (5bis)$$

Moreover, for an efficient point $(\mathbf{x}, \mathbf{y})$ we determine $r(\lambda)$ as

$$\Phi(\lambda \mathbf{x}, \lambda^{r(\lambda)} \mathbf{y}) = 0, \qquad (6)$$

and define

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

Observe that condition (6) amounts to the consideration of a point along the boundary of *T* generated by a contraction/expansion $\lambda > 0$ of the input vector of an efficient point $(\mathbf{x}, \mathbf{y})$ such that $\Phi(\mathbf{x}, \mathbf{y}) = 0$: i.e., $r(\lambda)$ is endogenously determined by the efficient production frontier.

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

Proposition 1: Under assumptions (A.1)-(A.4), at a cost-efficient point $(\mathbf{x}^*, \mathbf{y}) \in T$ the following equalities hold $r = S = \hat{S}$.

Proposition 1 reformulates Corollary 2 of Panzar and Willig (1977) by using the transformation function, typical for the neoclassical approach, in conjunction with the production frontier determined by (6), a concept mainly developed in the axiomatic approach. An analogous proposition is offered in Färe and Primont (1995, p. 53), but our proof differs because of the use of

Panzar and Willig's transformation function and of the ensuing possibility of establishing an immediate link between scale elasticities and RAP in a smooth technology. Following up on some important economic contributions like those of Menger (1954) and Frisch (1965) dealing with average productivity in a multiple inputs-single output setting, Banker (1984) extends this concept to multiple outputs production. More recently, Podinovski (2004, section 11.2) has introduced the term RAP to denote the ratio of the proportional increase in outputs to the proportional increase in inputs along the efficient frontier. Therefore, from (6) the RAP function can be easily obtained as $\pi(\lambda) = \lambda^{[r(\lambda)-1]}$. This allows to determine its derivative with respect to λ as:

$$\frac{\partial \pi}{\partial \lambda} = \lambda^{[r(\lambda)-1]}\left[\frac{\partial r}{\partial \lambda}\ln\lambda + \frac{\left(r(\lambda)-1\right)}{\lambda}\right]. \tag{8}$$

Expression (8) evaluated at $\lambda = 1$ yields:

$$\lim_{\lambda \to 1} \frac{\partial \pi}{\partial \lambda} = r - 1 . \tag{9}$$

Expression (9) intuitively shows that the behavior of the RAP function corresponds to the behavior of $r(\lambda)$ over the λ domain. Recent research shows that in *T* the RAP function is in general not concave (see Cesaroni et al. 2017 and Podinovski 2004) and that even in T_C the RAP function is only quasi-concave, as proven by Podinovski (2004, p. 246). Consequently, over some interval in both *T* and T_C, $r(\lambda)$ may be increasing: in these cases the use of the operator *sup* in the definition of scale elasticities (4) and (5) is not legitimate, because of the implied reference to a frontier point which is not coincident with the one under examination. To the best of our knowledge, this conclusion - which motivates the necessity of reformulating both duality result of Panzar and Willig

(1977) - is new to the literature, as witnessed by the use of the *sup* operator by Färe et al. (1988, p. 724 and p. 728) in the definition of scale elasticities.[4]

The conclusions so far leave us with two open problems. First, how to find conditions under which $r(\lambda)$ is monotonically decreasing and the *sup* definitions can be used? Second, given the former, how to determine a relationship between primal and dual elasticities in a non-smooth technology? As we see in the remainder of this contribution, the analysis of cost and ray average cost functions in a convex technology offers adequate solutions to these issues.

3. Ray Average Cost

The concept of RAC as a tool to investigate scale economies in multiple output technologies dates back to the seminal contributions of Baumol (1977), Baumol and Fisher (1978) and Baumol et al. (1982) in a setting where inefficiency in the production 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, p. 308) are the first to explicitly define the RAC function in the context of VRS production technologies with inefficiency and to point out that the concept is strictly connected to the evaluation of scale inefficiency with respect to the constant returns to scale (CRS) cost function.

Contrary to the more traditional approach in economics based on average productivity, this strand determines SE by means of the behavior of the cost function in response to equiproportional variations in outputs at given input prices (see, e.g., Baumol 1977, beginning of p. 811, and proposition 1, p. 812). Constant SE arise when average cost remains stationary at a minimum level.

[4] At p. 724, Färe et al. (1988) use the *sup* operator, while at p. 728 they literally state that the validity of their analysis does not require that the scale elasticity *E* declines monotonically.

Otherwise, we have increasing or decreasing SE depending on whether the average cost is decreasing with an increase or a decrease in the output's scale size, respectively (see definition 5 in Panzar and Willig 1977; cf. also Sueyoshi 1999 - section 4.2.1, and Tone and Sahoo 2003 - section 4.2). In any case, to the best of our knowledge no contribution addresses the issue of the RAC shape, which has so far been conceived as an assumption rather than a consequence of the form of the technology (see, e.g., Baumol and Fisher 1978 p. 460, Chavas and Cox 1999 p. 307, Baumol et al. 1982 p. 257).

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*.

3.1 Preliminaries: RAC Function, Optimal Scale Sizes and Scale Efficiency

Adopting Baumol's definition (see Baumol 1977, p.811; Baumol et al. 1982, pp. 48-49), the RAC function of an output vector $\mathbf{y}$ is

$$RAC(t,\mathbf{y}) = \frac{C(t\mathbf{y})}{t},\ t>0, \qquad (10)$$

where for notational convenience we have suppressed the input price vector $\mathbf{w}$ -which we assume as given- from the cost function at numerator and $C(t\mathbf{y}) = \mathbf{w}\mathbf{x}^*$ with $(\mathbf{x}^*(\mathbf{w},t\mathbf{y}),t\mathbf{y}) \in T$.

Chavas and Cox (1999, pp. 307-308) establish that if $\mathbf{y} \geq \mathbf{0}$, then a minimum of (10) with respect to *t* (RAC^*) exists under mild conditions which our technology satisfies, and that it is equal to the minimum cost of $\mathbf{y}$ in T^{CRS}. In other words, at given input prices the minimum RAC of an output mix in a VRS production technology corresponds to the value determined by the CRS cost function.

Following Cesaroni and Giovannola (2015, p. 123), a radial scaling factor between the output vectors of two different production possibilities belonging to T -the one under examination j and a generic reference unit h- can be derived from this operational definition[5]:

$$\gamma_{j,h} = \max_{r}\left\{\frac{y_{rj}}{y_{rh}}\right\}, \text{ where } \gamma_{j,h} \in (0,\infty] \tag{11}$$

Then, by putting $\frac{1}{t} = \gamma_{j,h}$ in (10), we can interpret its *argmin* as the production possibility that determines RAC^*, which we term optimal scale size (OSS): $(\mathbf{x}_o, \mathbf{y}_o) \in T$.

Therefore, if we define an efficiency measure as the ratio

$$R_j^* = \frac{RAC^*}{C(\mathbf{y}_j)}, \tag{12}$$

then it is evident that this efficiency measure equals one if and only if the output mix j under examination is an OSS[6], otherwise it is less than one. Moreover, R_j^* can be readily interpreted as the cost measure of scale efficiency of point $(\mathbf{x}_j^*(\mathbf{w}_j, \mathbf{y}_j), \mathbf{y}_j)$[7] (see Färe and Grosskopf 1985). Furthermore, the associated $\gamma_{j,o}$ is a straightforward indicator of global SE, that is the direction in T -increasing ($\gamma_{j,o} < 1$), decreasing ($\gamma_{j,o} > 1$), or sub-constant (both $\gamma_{j,o} < 1$ and $\gamma_{j,o'} > 1$)- along which outputs can be moved to achieve the minimum RAC (see Cesaroni et al. 2017b, p.1446).

[5] We consider the set of affinely extended real numbers.

[6] See Proposition 3 in Cesaroni and Giovannola (2015, p. 124), which states that ratio (12) equals unity for the output mix of an Oss.

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

3.2 RAC function within the Range between Two Optimal Scale Sizes

The former sub-section allows to address two important issues regarding the multiplicity of OSS points (RAC minima) and the behavior of RAC in the range comprised between two OSS points. To this end, we present a useful lemma.

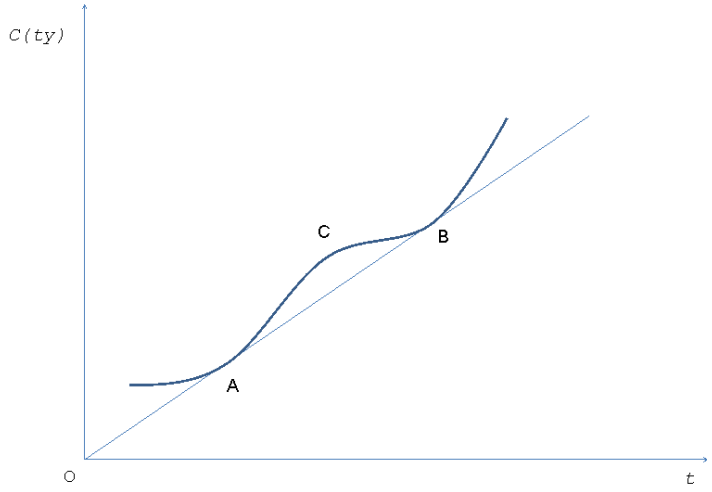
Lemma 1: The OSS of a cost-efficient point $(\mathbf{x}_j^*, \mathbf{y}_j) \in T$ is unique up to a positive multiplicative constant.

Lemma 1 reveals that multiple solutions to the minimization of RAC can occur if and only if exact proportional replicas of an OSS $(\mathbf{x}_o, \mathbf{y}_o)$ are present in *T*. This lemma is useful in ascertaining the behavior of $C(t\mathbf{y}_j)$ as a function of *t* in the interval comprised between two OSS. In fact, if the cost-efficient point *j* has two OSS, *o* and *o'* with $\gamma_{j,o} > 1 \, \gamma_{j,o'} < 1$ respectively, this amounts to examining the behavior of *C* in $\left[\underline{t}, \bar{t}\right]$, where $\underline{t} = \frac{1}{\gamma_{j,o}}$ and $\bar{t} = \frac{1}{\gamma_{j,o'}}$.

Theorem 1: In T_C, for a cost-efficient output mix $C(t\mathbf{y}_j) = RAC^* \;\; \forall t \in \left[\underline{t}, \bar{t}\right]$.

In graphical terms we point out that Theorem 1 implies that for a convex technology the behavior of the VRS cost function depicted in Figure 1 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., RAC^*. By contrast, Figure 1 may describe the shape of the VRS cost function in a nonconvex technology such as *T*. Note that, for given input prices, the RAC function is given in this figure as the ratio $C(ty)/t$, that is as the slope of the ray joining the origin and a point on $C(ty)$.

Figure 1. Nonconvex cost function with multiple RAC minima



3.3 RAC Function Outside the Range Including an Optimal Scale Size

Theorem 1 establishes that in T_C the RAC function is stationary (i.e., $R_j^*(\gamma_{j,o}) = 1$) in the interval between two OSS. Therefore, given this result, we only need to determine the behavior outside the above-mentioned range, that is to say the case where the output mix $\mathbf{y}_j$ is expanding/contracting towards the level determined by an OSS $(\mathbf{x}_o, \mathbf{y}_o)$, $\frac{1}{\gamma_{j,o}} \cdot \mathbf{y}_j$ with $\gamma_{j,o} \neq 1$.

To this purpose, we can express any output level $\mathbf{y}' \in \left[\mathbf{y}_j, \frac{1}{\gamma_{j,o}} \cdot \mathbf{y}_j \right)$ as $\mathbf{y}' \equiv (1-\alpha) \cdot \mathbf{y}_j + \alpha \cdot (\frac{1}{\gamma_{j,o}} \mathbf{y}_j)$, where $\alpha \in [0,1)$ is a continuous control-parameter.[8] The RAC of output level $\mathbf{y}'$ -generated by $\mathbf{y}_j$ - can thus be expressed as

[8] For the ease of exposition, we are assuming that $\gamma_{j,o} < 1$, but the same representation obviously holds for the opposite case $\gamma_{j,o} > 1$.

$$\frac{\mathbf{w}_j \mathbf{x}_j^*(\alpha)}{t}, \tag{13}$$

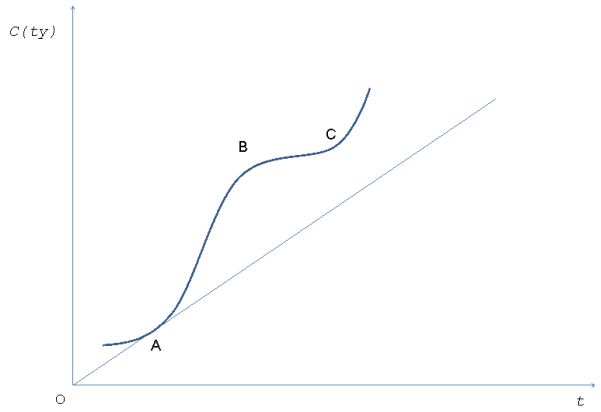
where $t \equiv \frac{\mathbf{y}'}{\overline{\mathbf{y}}_j}$ and $\mathbf{w}_j \mathbf{x}_j^*(\alpha) = C(\mathbf{y}')$. We can now point out that the problem under discussion may be solved by analyzing the derivative of expression (13) with respect to α. This leads to the following key result.

Theorem 2: In T_C, $RAC(\mathbf{y}_j)$ -where $\mathbf{y}_j \geq 0$ is a cost-efficient but not scale-efficient output mix- is monotonically decreasing and convex in either an expansion or a reduction to the output level of an OSS.

The algebraic proof employs a Shephard (1970) result on the cost function to determine the sign of the first and second derivative of (13) with respect to α at $\alpha = 0$. Recall that Shephard's (1970) result can be used to justify an empirical property of the cost function in the outputs: the cost function is convex (nonconvex) in the outputs when the production technology is convex (nonconvex). As long as $\mathbf{w}_j \mathbf{x}_j^*(\alpha)$ and *t* are positive -i.e., a minimum RAC exists[9]- the proof holds for any convex technology including the case of nonparametric (polyhedral) production technologies (see Ray 2004). In graphical terms, we point out that Theorem 2 implies that the behavior depicted in Figure 2, which is possible in a nonconvex *T*, cannot occur for T_C: here, the curve-section to the right of point A is monotonically increasing and convex.

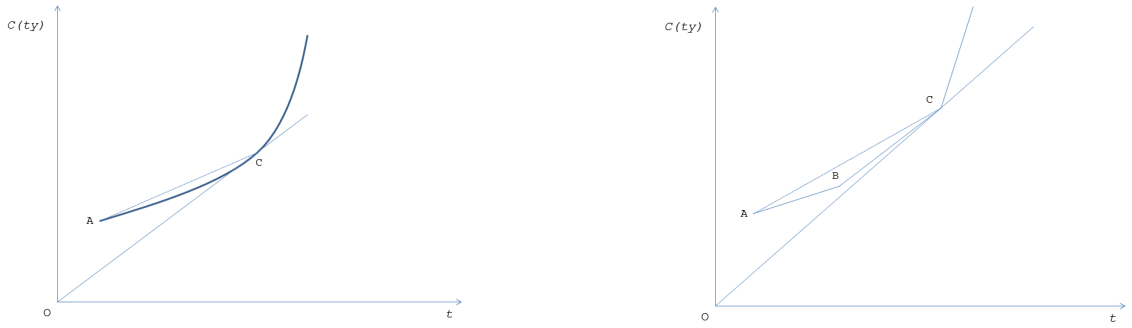
[9] See Chavas and Cox (1999, p. 307): footnote 18.

Figure 2. Nonconvex cost function with a unique RAC minimum



The cost function generated by a polyhedral technology corresponds to the case $\delta^{''}(0)=0$. Figure 3 illustrates the algebraic sign of $\delta^{'}(0)$ and $\delta^{''}(0)$ implied by the cost function of a convex technology to graphically illustrate Theorem 2. In Figure 3 (left), 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: $(1-\alpha)\cdot \mathbf{w}_j \mathbf{x}_j^* + \alpha \cdot \mathbf{w}_j \mathbf{x}_o$. Clearly, at point A: $\delta^{'}(0)<0$ and $\delta^{''}(0)>0$. Differently from the smooth case, in Figure 3 (right) we have two kinds of cost-scale inefficient points, i.e., point A and point B. For point A, we note that $\delta^{'}(0)<0$ and $\delta^{''}(0)=0$, while for point B we have $\delta^{'}(0)=0$ and $\delta^{''}(0)=0$, because in this last case the linear approximation of the cost function coincides with the line segment BC.

Fig. 3 Cost function of a smooth technology (left) and a polyhedral technology (right)



Observe that in Figure 3 (right), we only need to assume that the right derivatives $\delta'(0)$ and $\delta''(0)$ exist. Conversely, in the case of a contraction to the output level of an optimal scale size (i.e., starting from a point located above C), we only need to assume that left derivatives exist. In other words, we are not imposing differentiability of $\delta(.)$ and $\delta'(.)$ in the proof of Theorem 2.

We conclude this sub-section by remarking that Theorems 1 and 2 prove that in a convex production technology the RAC function is convex. In other words, RAC is 'U-shaped'. To the best of our knowledge, this is a finding that is absent from 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 form of the RAC function is supplied. Figure 4 illustrates this shape in the case of a differentiable *C* and multiple OSSs in $[t_1, t_2]$.

Figure 4. U-shape of RAC with multiple minima

4. Scale Efficiency, Primal and Dual Scale Elasticities

In this section the U-shape of RAC is used to derive some fundamental properties concerning the relations among cost scale efficiency, primal and dual scale elasticities.

4.1 Cost Scale Efficiency and Dual Scale Elasticity

The specific relationship between cost scale efficiency and dual scale elasticity in T_C is a result that is unavailable in the literature (cf., e.g., Färe et al. 1988, Färe and Primont 1995), as a consequence of the difficulties in the literature in determining the RAC shape.

The first step to get the needed result consists in the differentiation of RAC in expression (10) with respect to *t* evaluated at $t=1$ yielding:

$$\left.\frac{dRAC(.)}{dt}\right|_{t=1} = C(\mathbf{y})\cdot\left[\frac{1}{\hat{\mathrm{S}}}-1\right], \tag{19}$$

where $\hat{S}$ is the dual scale elasticity (5bis). At points where *C* is not differentiable, expression (19) takes two values determined by $\hat{S}^-$ and $\hat{S}^+$, respectively the left and right limit at $\mathbf{y}$ of (5bis).

As a second step, observe that the proof of Theorem 2 - part I - unambiguously determines the sign of (19) for cost scale inefficient points. Accordingly, we can formulate

Proposition 2: When *C* is everywhere differentiable, at a cost-efficient point $(\mathbf{x}^*, \mathbf{y}) \in T_C$ we have:

$$\begin{aligned} \hat{S} > 1 &\Leftrightarrow SE < 1 \text{ and GISE} \\ \hat{S} = 1 &\Leftrightarrow SE = 1 \text{ and GCSE} \\ \hat{S} < 1 &\Leftrightarrow SE < 1 \text{ and GDSE} \end{aligned} \qquad (20)$$

where *SE* is the scale efficiency measure (12) 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[10]of *t*. As a consequence $\hat{S}^- \geq \hat{S}^+$, which combined with Proposition 2 readily explains the following result:

Corollary 1

$$\begin{aligned} \hat{S}^- \geq \hat{S}^+ > 1 &\Leftrightarrow SE < 1 \text{ and GISE} \\ \hat{S}^+ \leq 1 \leq \hat{S}^- &\Leftrightarrow SE = 1 \text{ and GCSE} \\ \hat{S}^+ \leq \hat{S}^- < 1 &\Leftrightarrow SE < 1 \text{ and GDSE} \end{aligned} \qquad (20bis)$$

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., discontinuity of *C* at a single OSS,

[10] According to definition (5bis), the dual scale elasticity at a point corresponds to the cosine of the angle between the line tangent to the cost function at that point and the horizontal axis. From Figure 3, 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).

discontinuity of *C* at either the left or right end of the interval determined by multiple OSS points, respectively.

Proposition 2 and Corollary 1 conclude that in T_C both the scale-elasticity and the scale efficiency criterion provide the same indications for the SE regime. In other words, the result reached by 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. However, note that while Podinovski's conclusion -which is obtained under assumption (A.4)- can be derived for cost-efficient points by means of our Propositions 1 and 2, the reverse is not true. In addition, our Corollary 1 extends the result to the non-differentiable case. Finally, the convexity of *C* and the U-shape of RAC enable to reach some new significant properties (see Propositions 3 and 4 below) that cannot be obtained on the basis of the RAP function, which is only quasi-concave in T_C (see above, end of Section 2.2).

4.2 Primal and Dual Scale Elasticities

Finally, we are in a position to generalize what we have called Result 2 in 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 result:

Proposition 3: At a cost-efficient point $(\mathbf{x}^*, \mathbf{y}) \in T_C$ where differentiability of *C* does not hold, we have

$$\begin{aligned} S^- &\geq \hat{S}^- \\ S^+ &\leq \hat{S}^+ \end{aligned} \qquad (21)$$

Proposition 3 reveals that $S \leq \hat{S}$ (i.e., proposition 5 in Panzar and Willig 1977, p. 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.

5. Correspondence between Returns to Scale and Scale Economies Regimes

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 measures (see, e.g., Tone and Sahoo 2003). 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, a specific relationship must be worked out.

To this purpose, Corollary 1 and Proposition 3 can be employed to yield the result:

Proposition 4: At a cost-efficient point $(\mathbf{x}^*, \mathbf{y}) \in T_C$ where differentiability of C does not hold, we have

$$\begin{array}{rcl} \text{ISE} & \Rightarrow & \text{IRS or CRS} \\ \text{CSE} & \Rightarrow & \text{CRS} \\ \text{DSE} & \Rightarrow & \text{DRS or CRS} \end{array} \qquad (22)$$

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 $(\mathbf{x}^*, \mathbf{y}) \in T_C$ where differentiability of C does not hold, we have

$$\begin{array}{rcl} \text{IRS} & \Rightarrow & \text{ISE} \\ \text{CRS} & \Rightarrow & \text{ISE or CSE or DSE} \\ \text{DRS} & \Rightarrow & \text{DSE} \end{array} \qquad (22\text{bis})$$

Relationships (22) and (22bis) 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 Baumol's 1977 jargon), as those of Cesaroni and Giovannola (2015) and Zelenyuk (2014).

6. Applications of the Main Results

6.1 Global and Local Methods for Scale Economies in Nonparametric Convex Production Technologies

The method of Färe and Grosskopf (1985: p. 600) computes cost scale efficiency defined as the ratio between CRS and VRS cost efficiency scores of points lying on the frontier of a convex nonparametric technology[11] (see (12) above). If the point under examination is cost scale efficient, then it displays CRS. If the examined point is not cost scale efficient, then a third non-increasing returns to scale (NIRS) technology is employed to establish qualitative information on the nature of cost scale inefficiency: i.e., on the direction to the VRS optimal scale size whose projection determines the CRS cost efficiency score. To be precise, when NIRS and VRS cost efficiencies coincide, then scale inefficiency is due to DRS: i.e., the optimal scale size is lower than the current one. Otherwise, when NIRS and VRS cost efficiencies do not coincide, then scale inefficiency is due to IRS: i.e., the optimal scale size is greater than the current one. This method ignores the possibility of multiple CRS solutions. Following Podinovski (2004), we denote this classification as global because it is based on the absolute-minimum cost (i.e., the CRS cost) and it is determined by

[11] As far as the assumption of boundedness of the feasible set is concerned, we point out that this assumption is not violated as long as the observations (observed input-output combinations) are bounded. See (A.3).

a scale size which may be rather distant from the current scale examined. Note that this method does not provide quantitative information relating to the degree of scale economies, as commonly measured at the local level by the scale elasticity $\hat{S}$.

This scale elasticity approach is developed by Sueyoshi (1999) to ascertain the scale economies regime of frontier points in the convex nonparametric technology. In addition to a VRS cost-efficiency linear programming problem, a complex problem, which is the dual of the former (see Sueyoshi 1999, pp. 1599-1601), is set to compute the cost-scale elasticity $\hat{S}$ at the projection of an observation onto the efficient frontier. In analogy with the returns to scale analysis of Banker et al. (1984) and Banker and Thrall (1992), this elasticity can be interpreted as the intercept of the supporting hyperplane in the cost-outputs space. The cost-efficient point under examination exhibits increasing, decreasing and constant SE if $\hat{S}^- > \hat{S}^+ > 1$, $\hat{S}^+ < \hat{S}^- < 1$, $\hat{S}^+ \leq 1 \leq \hat{S}^-$ hold, respectively (see Sueyoshi 1999, p. 1603). In particular, we may remark that this method does not discuss the relationship between a solution to the CRS cost-efficiency problem and a VRS solution featuring local constant SE - i.e., $\hat{S}^+ \leq 1 \leq \hat{S}^-$ -, while also multiple solutions to the CRS problem are not taken into account (cfr., *a contrario*, Banker 1984, Banker and Thrall 1992 in a RTS setting).[12]

Complementing the RAP analysis of Podinovski (2004), Cesaroni et al. (2017b) show that in a nonconvex technology the multiplicity of CRS solutions and the non-monotonicity of the RAC function may lead to both an erroneous global classification and a non-coincidence of global and local scale economies classifications. These problems are illustrated in Figures 1 and 2, where for the ease of exposition a smooth technology is assumed.

[12] Note that Sueyoshi's (1999) "multiple solutions" are the multiple values of the scale elasticity at a corner point (vertex) of the frontier, and not the multiple CRS solutions to the scale-efficiency problem. The relationship between this kind of multiplicity and the multiplicity of solutions in a reference set is discussed in Sueyoshi and Sekitani (2007), but only for production-based returns to scale.

Figure 1 describes the hypothetical case of multiple CRS solutions and nonconvex behavior of RAC. The Färe and Grosskopf (1985) criterion classifies point C as global increasing SE, while it is in fact characterized by global sub-constant SE.[13]

Moreover, we note that also the Sueyoshi (1999) local criterion yields a misleading result: since $\hat{S} = 1$, point C is denoted as CSE instead of the global sub-constant SE characterization. In other words, the local indicator does not detect the direction to an OSS.

This kind of divergence between local and global SE indicators can also take the form illustrated in Figure 2. Here, each point along the curve-section BC operates under local increasing scale economies ($\hat{S} > 1$) while these points are actually in a DSE regime from the global point of view -given that point A is their optimal scale size.

For these reasons, it is important to integrate CRS multiple-solutions in a global classification method and to determine the form of the RAC curve to check the correspondence between local and global indicators in a convex technology. Our study has just shown that even in the presence of multiple CRS solutions and a non-differentiable cost function, which may occur in a convex technology, the U-shape of RAC prevents those incorrect classifications, so that local and global scale economies indicators coincide. In other words, we have proven that the Färe and Grosskopf (1985) and the Sueyoshi (1999) methods can also be employed as local and global methods, respectively, for the estimation of SE in convex technologies and even in the presence of multiple CRS solutions.

With respect to the Färe and Grosskopf (1985) method, we point out that our Proposition 2 and its proof show that there is no necessity to compute the cost efficiency score relative to the NIRS technology, because to achieve the desired global SE classification the computation of $\gamma_{j,o}$ in the VRS technology suffices.

[13] See the numerical example in Cesaroni et al. (2017b, p. 1444).

6.2 Correspondence of Returns to Scale and Scale Economies Regimes in Nonparametric Convex Production Technologies

To the best of our knowledge, the only analysis dealing with the relation between RTS and SE in multi-output production when the frontier is not differentiable is offered in Sueyoshi (1999, section 4.2.2) for polyhedral convex technologies. In this regard, when compared to our study, two remarks are in order.

First, Sueyoshi's analysis does not hold for convex technologies other than polyhedral ones, while its conclusions are derived exclusively from linear programming programs and not from the properties of the cost function *C*. Second, perhaps as a consequence of this last shortcoming, it can be shown that his specific conclusions are incorrect.

In fact, besides failing to distinguish between technical-efficient and cost-efficient points on the frontier of the technology -while only for the latter exact conclusions on the relation between primal and dual scale elasticities can be reached (see, e.g., Panzar and Willig 1977, Färe and Primont 1995)- Sueyoshi (1999, p. 1604) presents the following scheme:

(a) if one of the two scale elasticities indicates IRS[14], then the other may take either IRS or CRS;

(b) if one of the two is CRS, then the other may take any type of RTS;

(c) if one of the two is DRS, then the other may take either DRS or CRS.

Our relationship (21) and (21bis) can be used to check the validity of this scheme for a cost-efficient point. This yields the following remarks:

(a) is valid for the dual scale elasticity, but not for the primal scale elasticity;

[14] Sueyoshi (1999) uses the same notation for RTS and SE regimes.

(b) is not valid for the dual scale elasticity, while it is valid for the primal scale elasticity;

(c) is valid for the dual scale elasticity, but not for the primal scale elasticity.

Therefore, we can conclude that the Sueyoshi (1999) correspondence-scheme between RTS and SE is wrong, as illustrated in the empirical section below.

6.3 RAC and the Determination of the Optimal Number of Firms in Baumol and Fisher (1978)

In their discussion of the determination of the cost-minimizing number of firms in a multi-product industry, Baumol and Fisher (1978) make, among others, the assumption of a U-shaped RAC with a unique minimum (see, ibid., assumption 2, pp. 441-442). Ten Raa (1983) shows that this only assumption -enlarged to take account of multiple RAC minima- is sufficient for the existence of the desired upper and lower bounds on the optimal number of firms (see, ibid., Revision 1, fn. 2 and Theorem 1, pp. 214-215). Therefore, our analysis, which allows for multiple RAC minima, can be applied to show that in any convex production technology, independently from the differentiability of *C*, the conditions for a determinate cost-minimizing number of firms are indeed satisfied.

7. Empirical Illustration

The theoretical results achieved in the previous sections are illustrated in detail using the data set of Cesaroni et al. (2017b) for the case of a nonparametric convex technology. This data set is published as supplemental material of the article: it is freely accessible. It concerns a three inputs-two outputs production technology associated with the Italian local-public-transit sector, where each firm faces a specific input price vector.

Corollary 1 of Proposition 2 is shown in Table 1 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 1. OE is obtained by substituting the numerator of expression (12) with the optimal cost in the VRS technology. The *Rj* ratio is given by expression (12), while the gamma ratio $\gamma_{j,o}$ is defined by expression (11). These variables have been computed by means of the R-package Benchmarking. Cost scale elaticities $\hat{S}^-$ and $\hat{S}^+$ are computed by means of models (34) and (30) in Sueyoshi (1999). These computations have been performed in Maple.

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

	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

We remark that computations empirically confirm the coincidence of global and local scale economies indicator (GSE and LSE) in the case of a convex production technology at points where *C* is not differentiable, as established by Corollary 1.

Table 2 employs the results obtained from the application of Sueyoshi (1999) computation method regarding primal and dual scale elasticities, (S^-, S^+) and $(\hat{S}^-, \hat{S}^+)$ 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 have been obtained from models (33) and (14) in Sueyoshi (1999), while the dual scale elasticity is computed as indicated when discussing Table 1. Computations have been performed in Maple.

Table 2. Primal and dual scale elasticities, RTS and SE

	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

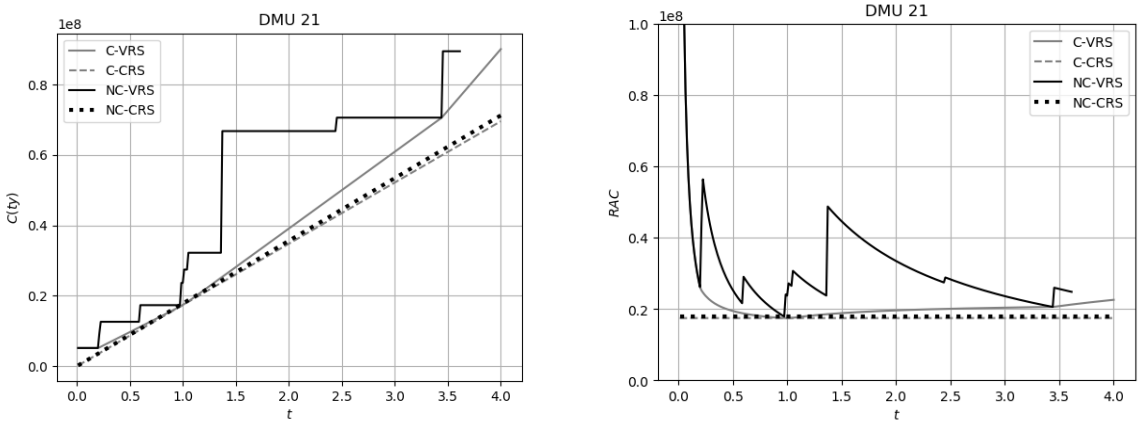
Table 2 makes evident that inequalities (21) of Proposition 3 empirically hold, but also that the correspondences (22) and (22bis) (associated to Proposition 4) are satisfied (see columns RTS and SE). We point out that the empirical validity of Proposition 4 confirms that Sueyoshi's (1999) interpretation of the relationship between RTS and SE classification is not appropriate. In fact, as

far as our remark (b) and (c) in Section 6.2 are concerned[15], note that each of the two constant SE observations (i.e., 9 and 32) features also constant RTS, while each of the decreasing RTS observations (i.e., 5, 22, 29, 39) is in a decreasing regime -not in a constant- also on the SE side.

Finally, Figure 5 illustrates the different shape that VRS and CRS cost and ray-average-cost functions assume in convex (C) and nonconvex (NC) nonparametric technologies (see Ray 2004, p. 57 and p. 137 resp.). This is being done with reference to the output mix of observation 21: Figure 5 (left) plots $C(ty)$ against t (just like in Figures 1 and 2), while Figure 3 (right) plots RAC against t. Starting from positive values of t near to 0 and expanding the output mix towards 1, the convex VRS (C-VRS) technology clearly exhibits a U-shaped RAC (i.e., the ratio $C(ty)/t$ is a convex function). By contrast, in addition to the discontinuities generated by the horizontal segments of its cost function, the non-convex VRS (NC-VRS) technology presents a RAC which is not U-shaped, because it is simultaneously increasing/decreasing both for $t<1$ and $t>1$ (i.e., the ratio $C(ty)/t$ is nonconvex). Thus, not only the cost function behaves differently depending on the convexity or nonconvexity of the underlying technology, but also the RAC function shares this same property: RAC is convex (nonconvex) in the outputs when the production technology is convex (nonconvex).

[15] The absence of DMUs with increasing returns to scale does not allow to show the validity of our remark (a) in section 6.2.

Fig. 5 Cost function C(ty) (left) and RAC (right) for DMU 21



8. Conclusions

This contribution advances the literature on scale economies in convex production technologies in several ways. First, it determines the U-shape of the RAC function and extends the coincidence of local and global returns to scale from production to cost analysis in 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 quantitative information, by means of the $\gamma_{j,o}$ coefficient, that is otherwise not available. Third, besides showing that a required monotonicity condition 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.

References

Banker, R.D. (1984) Estimating most productive scale size using Data Envelopment Analysis, *European Journal of Operational Research 17,* 35-44.

Banker, R.D., Chang, H., Cooper, W.W. (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., Cooper, W.W. (1984) Models for the estimation of technical and scale inefficiencies in Data Envelopment Analysis, *Management Science 30*, 1078-1092.

Banker, R.D., Thrall, R.M. (1992) Estimation of returns to scale using Data Envelopment Analysis, *European Journal of Operational Research 62*, 74-84.

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., Fisher D. (1978) Cost-minimizing number of firms and determination of industry structure, *Quarterly Journal of Economics* 92, 439-468.

Baumol, W.J., Panzar, J.C., Willig, R.D. (1982) *Contestable markets and the theory of industry structure*. New York: Harcourt Brace Jovanovich.

Chavas, J.P., Cox, T.L. (1999) A generalized distance function and the analysis of production efficiency, *Southern Economic Journal 66*, 294-318.

Cesaroni, G., Giovannola, D. (2015) Average-cost efficiency and optimal scale sizes in non-parametric analysis, *European Journal of Operational Research 242*, 121-133.

Cesaroni, G., Kerstens, K., Van De Woestyne, I. (2017a) Global and local scale characteristics in convex and nonconvex nonparametric technologies: A first empirical exploration, *European Journal of Operational Research 259*, 576-586.

Cesaroni, G., Kerstens, K., Van De Woestyne, I. (2017b) Estimating scale economies in non-convex production models, *Journal of the Operational Research Society 68*, 1442-1451.

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

Färe, R., Grosskopf, S., Lovell, C.A.K. (1988) Scale elasticity and scale efficiency, *Journal of Institutional and Theoretical Economics 144*, 721-729.

Färe, R., Grosskopf, S. (1985) A nonparametric cost approach to scale efficiency, *Scandinavian Journal of Economics 87*, 594-604.

Førsund, F., Hjalmarsson, L. (2004) Are all scales optimal in DEA? Theory and empirical evidence, *Journal of Productivity Analysis 21*, 25-48.

Førsund, F., Kittelsen S., Krivonozhko V. (2009) Farrell revisited-Visualizing properties of DEA production frontiers, *Journal of the Operational Research Society 60*, 1535-1545.

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.

Menger, K. (1954) The logic of the laws of return: A study in meta-economics, in Morgenstern O., Ed., *Economic activity analysis* Part III, New York: Wiley.

Panzar, W., Willig, R.D. (1977) Economies of scale in multi-output production, *Quarterly Journal of Economics 91*, 481-493.

Podinovski, V. (2004) Efficiency and global scale characteristics on the "No free lunch" assumption only, *Journal of Productivity Analysis 22*, 227-257.

Ray, S.C. (2004) *Data envelopment analysis: Theory and techniques for economics and operations research*, Cambridge: Cambridge University Press.

Ray, S. (2015) Nonparametric measures of scale economies and capacity utilization: An application to U.S. manufacturing, *European Journal of Operational Research 245*, 602-611.

Shephard, R.W. (1970) *Theory of cost and production functions.* Princeton: Princeton University Press.

Sueyoshi, T. (1999) DEA duality on returns to scale in production and cost analyses: An occurrence of multiple solutions and differences between production-based and cost-based RTS estimates, *Management Science 45*, 1593-1608.

Sueyoshi, T., Sekitani K. (2007) The measurement of returns to scale under a simultaneous occurrence of multiple solutions in a reference set and a supporting hyperplane, *European Journal of Operational Research 181*, 549-570.

Ten Raa, T. (1983) On the cost-minimizing number of firms, *Economics Letters 12*, 213-218.

Tone, K., Sahoo, B.K. (2003) Scale indivisibilities and production function in Data Envelopment Analysis, *International Journal of Production Economics 84*, 165-192.

Zelenyuk, V. (2014) Scale efficiency and homotheticity: equivalence of primal and dual measures, *Journal of Productivity Analysis 42*, 15-24.

Appendices (Online supplement)

Appendix 1: Result 1) of Panzar and Willig (1977)

As pointed out in Section 2.1, the duality result established by Corollary 2 in Panzar and Willig (1997, p. 491) is not correct, because of a specific inconsistency that affects the proof of Theorem 1 (see, ibid., pp. 489-490).

Albeit authors state that under their axioms "efficient production exists" (see, ibid., p.487) they do not employ the notion of efficient frontier in the proof in question. This notion, represented by equality (6), when considered in the proof yields $H'(1)=0$ and $\varepsilon = r$; as a consequence, ε is not an upper bound on R while no contradiction arises from $S \geq \varepsilon$, therefore Corollary 1 is wrong because it holds with the inequality sign, i.e. $S \geq -\dfrac{\sum_i x_i^* \partial\Phi/\partial x_i}{\sum_j y_j \partial\Phi/\partial y_j}$. Now, note that this inequality sign combined with Theorem 2 implies $S \geq \hat{S}$, that is Corollary 2 is wrong (see, ibid., pp. 490-491).

Appendix 2: Proofs

Proof of Proposition 1: At a cost-efficient point $\mathbf{x}^*(\mathbf{w},\mathbf{y}) \in T$, $\mathrm{S} = \hat{\mathrm{S}}$ is proven as equality (5) in Panzar and Willig (1977, pp. 488-489). Furthermore, given expression $\Phi(\lambda \mathbf{x}^*, \lambda^{r(\lambda)}\mathbf{y}) = 0$ and assumption (A.5), $\dfrac{d\Phi(\lambda \mathbf{x}^*, \lambda^{r(\lambda)}\mathbf{y})}{d\lambda} = 0$ yields $\dfrac{d\Phi(\lambda \mathbf{x}^*, \lambda^{r(\lambda)}\mathbf{y})}{d\lambda} = \sum_i x_i^* \partial\Phi/\partial x_i + \lambda^{r(\lambda)}\left(\partial r/\partial\lambda \cdot \ln\lambda + r(\lambda)/\lambda\right)\cdot \sum_j y_j \partial\Phi/\partial y_j = 0$. Finally, evaluating this expression at $\lambda = 1$ gives $r = \mathrm{S}$. Q.E.D.

Proof of Lemma 1: Suppose that point *j* has two different optimal scale sizes, *o* and *o'*. Compare first *o'* with *o*. From a minimum RAC property of an optimal scale size (see Proposition 3 in Cesaroni and Giovannola 2015) we have $\frac{\mathbf{w}_j\mathbf{x}_o}{\mathbf{w}_j\mathbf{x}_{o'}}\cdot\gamma_{o',o}\geq 1$, where $\gamma_{o',o}=\max_r\left\{\frac{y_{ro'}}{y_{ro}}\right\}$. In addition, the minimization of point *j*'s RAC implies $\mathbf{w}_j\mathbf{x}_{o'}\cdot\gamma_{j,o'}=\mathbf{w}_j\mathbf{x}_o\cdot\gamma_{j,o}$, from which we have (a1) $\frac{\mathbf{w}_j\mathbf{x}_o}{\mathbf{w}_j\mathbf{x}_{o'}}=\frac{\gamma_{j,o'}}{\gamma_{j,o}}$. We can then express the former inequality as (a2) $\frac{\gamma_{j,o'}}{\gamma_{j,o}}\cdot\max_r\left\{\frac{y_{ro'}}{y_{ro}}\right\}\geq 1$. Moreover, the property of being a most productive scale size implies that the RAP of *o'* with respect to *o* is equal or greater than one, therefore we have (a3) $\frac{\gamma_{j,o}}{\gamma_{j,o'}}\cdot\frac{1}{\max_r\left\{\frac{y_{ro'}}{y_{ro}}\right\}}\geq 1$ (see Proposition 4 and expression (10) in Cesaroni and Giovannola 2015). Now, note that (a3) is the reciprocal of (a2): both inequalities can hold if and only if the equality sign applies. Selecting inequality (a2), we must then have (a4) $\frac{\gamma_{j,o'}}{\gamma_{j,o}}\cdot\max_r\left\{\frac{y_{ro'}}{y_{ro}}\right\}=1$. The same reasoning applies in the comparison between *o* and *o'*, so that (a5) $\frac{\gamma_{j,o}}{\gamma_{j,o'}}\cdot\max_r\left\{\frac{y_{ro}}{y_{ro'}}\right\}=1$ follows as well. Employing (a4), (a5) can be rewritten as $\max_r\left\{\frac{y_{ro'}}{y_{ro}}\right\}\cdot\max_r\left\{\frac{y_{ro}}{y_{ro'}}\right\}=1$. Because of $\frac{1}{\max_r\left\{\frac{y_{ro}}{y_{ro'}}\right\}}\equiv\min_r\left\{\frac{y_{ro'}}{y_{ro}}\right\}$, (a5) can be finally read as $\max_r\left\{\frac{y_{ro'}}{y_{ro}}\right\}=\min_r\left\{\frac{y_{ro'}}{y_{ro}}\right\}$, which amounts to establishing that output vectors $\mathbf{y}_{o'}$ and $\mathbf{y}_o$ are equal up to a multiplicative constant, $\mathbf{y}_{o'}=\alpha\cdot\mathbf{y}_o$. Equality (a4) can then be used to show that $\alpha=\frac{\gamma_{j,o}}{\gamma_{j,o'}}$, while $\mathbf{x}_{o'}=\alpha\cdot\mathbf{x}_o$ because of the identity of input prices in (a1). Q.E.D.

Proof of Theorem 1: Suppose not, then $R_j^* < 1$. Lemma 1 has proven that each optimal scale size is a proportional replica of the other. This implies that in T_C, by taking a suitable linear convex combination of *o* and *o'*, an OSS $\bar{o}$ can always be obtained such that $\gamma_{j,\bar{o}} = 1$. Consequently $RAC^* = C(\mathbf{y}_{\bar{o}})$, while $\mathbf{y}_{\bar{o}} \geq \mathbf{y}_j$ and the cost-efficiency of j imply $C(\mathbf{y}_{\bar{o}}) \geq C(\mathbf{y}_j)$. Therefore, $RAC^* = C(\mathbf{y}_{\bar{o}}) \geq C(\mathbf{y}_j)$ and $R_j^* \geq 1$, which contradicts the assumption $R_j^* < 1$. Q.E.D.

Proof of Theorem 2: Shephard (1970) proves that the cost function is convex in the outputs if the production technology is convex (see, ibid., the definition of graph, pp. 180-181, and proposition 72 - property Q.11, p. 227). This allows to get equality $\mathbf{w}_j \mathbf{x}_j^*(\alpha) = (1-\alpha) \cdot \mathbf{w}_j \mathbf{x}_j^* + \alpha \cdot \mathbf{w}_j \mathbf{x}_o + \delta(\alpha)$, with $0 \leq \alpha < 1$ and $\delta(\alpha) \leq 0$. Thus, we can express the RAC (13) as

$$\frac{\mathbf{w}_j \mathbf{x}_j^* + \alpha \cdot (\mathbf{w}_j \mathbf{x}_o - \mathbf{w}_j \mathbf{x}_j^*) + \delta(\alpha)}{1 + \alpha \cdot (\frac{1}{\gamma_{j,o}} - 1)} \qquad \text{(13bis)}$$

The proof is made up of two parts.

Part 1: Monotonicity

By differentiating (13bis) with respect to α and rearranging, we obtain the final expression of the derivative as

$$\frac{(\mathbf{w}_j \mathbf{x}_o - \frac{1}{\gamma_{j,o}} \cdot \mathbf{w}_j \mathbf{x}_j^*) + [\delta'(\alpha) + (\frac{1}{\gamma_{j,o}} - 1) \cdot (\alpha \delta'(\alpha) - \delta(\alpha))]}{\left[1 + \alpha \cdot (\frac{1}{\gamma_{j,o}} - 1)\right]^2} \qquad \text{(14bis)}$$

where $\delta'(\alpha)$ denotes the derivative of δ with respect to α.

Being $\delta(0) = 0$, (14bis) evaluated at $\alpha = 0$ reduces to

$$(\mathbf{w}_j\mathbf{x}_o - \frac{1}{\gamma_{j,o}} \cdot \mathbf{w}_j\mathbf{x}_j^*) + \delta^{'}(0) \qquad (15bis)$$

This expression is negative because of: a) $(\mathbf{w}_j\mathbf{x}_o - \frac{1}{\gamma_{j,o}} \cdot \mathbf{w}_j\mathbf{x}_j^*) < 0$, where this last inequality comes from the assumption of $(\mathbf{x}_o, \mathbf{y}_o)$ being the optimal scale size of $(\overline{\mathbf{x}}_j, \overline{\mathbf{y}}_j)$: $\gamma_{j,o\cdot} \cdot \mathbf{w}_j\mathbf{x}_o < \mathbf{w}_j\mathbf{x}_j^*$, and from the positivity of $\gamma_{j,o}$; and b) $\delta^{'}(0) \leq 0$ - resulting from the convexity of the cost function in the outputs.

The derivative of the RAC function is negative at $\alpha = 0$ - i.e. at the initially selected value $\mathbf{y}_j$ - but this value can be arbitrarily close to or distant from $\frac{1}{\gamma_{j,o}} \cdot \mathbf{y}_j$ because $\gamma_{j,o}$ can take any positive value, therefore negativity holds for each $\mathbf{y}^{'} \in \left[\mathbf{y}_j, \frac{1}{\gamma_{j,o}} \cdot \mathbf{y}_j\right)$. Note moreover that this conclusion is independent from $\gamma_{j,o} < 1$: it identically holds for $\gamma_{j,o} > 1$, i.e., for each $\mathbf{y}^{'} \in \left(\frac{1}{\gamma_{j,o}} \cdot \mathbf{y}_j, \mathbf{y}_j\right]$.

Part 2: Convexity

The second derivative of the RAC with respect to α can be obtained from the differentiation of (14). After some manipulations, we get its final expression as

$$\frac{\delta^{''}(\alpha)}{\left[1 + \alpha \cdot (\frac{1}{\gamma_{j,o}} - 1)\right]} - \frac{2 \cdot num \cdot (\frac{1}{\gamma_{j,o}} - 1)}{\left[1 + \alpha \cdot (\frac{1}{\gamma_{j,o}} - 1)\right]^3} \qquad (16bis)$$

where $\delta^{''}(\alpha)$ indicates the second derivative with respect to α and *num* is the numerator of (14bis). Now, at $\alpha = 0$, (16bis) reduces to

$$\delta^{''}(0) - 2 \cdot num(0) \cdot (\frac{1}{\gamma_{j,o}} - 1) \qquad (17bis)$$

where *num(0)* is expression (15bis), which has a negative sign. Moreover, note that convexity of the cost function in the outputs implies $\delta^{''}(0) \geq 0$ and $\delta^{''}(0) \leq 0$ for $\gamma_{j,o} < 1$ and $\gamma_{j,o} > 1$, respectively.

Therefore, (17bis) is positive for $\gamma_{j,o} < 1$, and negative for $\gamma_{j,o} > 1$. This implies that the RAC function is convex in both an expansion and a contraction to an optimal scale size. Q.E.D.

Proof of Proposition 2: Observe that $\hat{S} > 1$ and $\hat{S} < 1$ respectively yield a negative and positive value of (19); given Theorems 1 and 2, the negative value implies $SE < 1$ and GISE - i.e. $\gamma_{j,o} < 1$ -, while the positive implies $SE < 1$ and GDSE - i.e. $\gamma_{j,o} > 1$. Moreover, the reverse of these two implications is immediately ensured by Theorem 2. As for $\hat{S} = 1$, note that it implies a minimum RAC, but Theorem 1 shows that this is equivalent to $SE = 1$ and GCSE - i.e. $\gamma_{j,o} = 1$, while the reverse implication follows from the definition of an optimal scale size and the differentiability of *C*. Q.E.D.

Proof of Proposition 3: Suppose that A.4 holds for $\Phi(\lambda \mathbf{x}^*, \lambda^{r(\lambda)} \mathbf{y}) = 0$ in a small symmetrical neighborhood of $\lambda = 1$, but not at $\lambda = 1$. In this neighbourhood, the convexity of *C* makes $\hat{S}$ to be a non-increasing function of *t* while Proposition 1 establishes that $r = \hat{S}$, these imply that also r is a non-increasing function of *t*. Therefore, we have

$$S^{-} = inf\left\{ r \middle| \exists \delta < 1 \text{ such that } (\lambda \mathbf{x}, \lambda^{r} \mathbf{y}) \in T \text{ for } \delta \leq \lambda \leq 1 \right\}$$

$$S^{+} = sup\left\{ r \middle| \exists \delta > 1 \text{ such that } (\lambda \mathbf{x}, \lambda^{r} \mathbf{y}) \in T \text{ for } 1 \leq \lambda \leq \delta \right\}$$

$$\hat{S}^{-} = inf\left\{ r \middle| \exists \delta < 1 \text{ such that } C(\lambda \mathbf{y}, \mathbf{w}) \leq \lambda^{1/r} C(\mathbf{y}, \mathbf{w}) \text{ for } \delta \leq \lambda \leq 1 \right\},$$

$$\hat{S}^{+} = sup\left\{ r \middle| \exists \delta > 1 \text{ such that } C(\lambda \mathbf{y}, \mathbf{w}) \leq \lambda^{1/r} C(\mathbf{y}, \mathbf{w}) \text{ for } 1 \leq \lambda \leq \delta \right\}.$$

Now, if we denote by R and $\hat{R}$ the set of values of r such that, respectively, the second and the fourth equality hold, it easy to check that $R \subseteq \hat{R}$ (see Panzar and Willig 1977, p. 491), so that $S^{+} \leq \hat{S}^{+}$. Note that $R \subseteq \hat{R}$ also holds when these sets are being referred to the first and third equality, respectively, therefore $S^{-} \geq \hat{S}^{-}$. Q.E.D.

Proof of Proposition 4: When (A.4) does not hold, the same relations between values of left and right scale-elasticity shown in (20bis) also apply to S^{-} and S^{+} in the definition of IRS, CRS and DRS respectively (see, e.g., definition 8 in Banker and Thrall 1992, p 79). Then, the implications in (22) easily follow from the three cases of $\hat{S}^{-}$ and $\hat{S}^{+}$ in (20bis) combined with inequalities (21). Q.E.D.