Journal of the Operational Research Society (2017) 68, 1442–1451 © 2017 The Operational Research Society. All rights reserved. 0160-5682/17

www.palgrave.com/journals

Estimating scale economies in non-convex production models

Giovanni Cesaroni[1]*, Kristiaan Kerstens[2] and Ignace Van de Woestyne[3]

[1]Department for Economic Policy, Prime Minister's Office, Via della Mercede 9, 00187 Rome, Italy; [2]CNRS-LEM (UMR 9221), IESEG School of Management, 3 rue de la Digue, 59000 Lille, France; and [3]Research Unit MEES, KU Leuven, Warmoesberg 26, 1000 Brussels, Belgium

The literature on nonparametric frontier technologies lacks a method for the measurement of scale economies in non-convex settings. This paper proposes a general procedure which is based on the minimization of the ray average cost and requires the solution of a single programming problem. Our approach allows for multiple optima to introduce the case of global sub-constant scale economies, and it also permits the estimation of scale economies at a local level. The empirical application investigates the role of replicability and the relationship between global and local indicators. It also points out the managerial implications for companies operating in the Italian public transit industry.

Journal of the Operational Research Society (2017) **68(11),** 1442–1451. doi:10.1057/s41274-016-0162-7; published online 8 February 2017

Keywords: Free Disposal Hull; returns to scale; scale economies

1. Introduction

The issue of the measurement of scale economies in nonparametric production models has so far attracted rather little analytical interest when compared to the vast literature on "production-based returns to scale"[1] (e.g., Banker, 1984; Banker *et al*, 1984; Banker and Thrall, 1992; Banker *et al*, 1996; Färe *et al*, 1983, 1985; Kerstens and Vanden Eeckaut, 1999). The latter contributions invariably focus on the maximization of ray average productivity of a decision making unit (DMU) without considering the possible allocative inefficiency of its input mix. Given the importance of allocative considerations in the evaluation of costs and its specific bearing on the notion of optimal scale size, this relative neglect is a bit strange. This same focus is actually found in earlier seminal studies on returns to scale and economies of scale in multiple output technologies ignoring inefficiency (Panzar and Willig, 1977; Baumol *et al*, 1982). In this sense, Tone and Sahoo's (2003) remark that "scale in all its definitions warrants the input mix and output mix to remain constant" is pertinent.

The available methods for the estimation of "cost-based returns to scale" with a variable input mix are at present limited to those developed by Färe and Grosskopf (1985) and Sueyoshi (1999) for convex production models. However, these methods are unsuitable for application to a non-convex technology (e.g., the Free Disposal Hull (FDH) of Deprins *et al*, 1984; Tulkens, 1993). On the one hand, Sueyoshi's measure of cost-scale elasticity provides local quantitative information on the degree of scale economies, but this cannot be defined due to the non-differentiability of the FDH technology (see Sueyoshi, 1999, p. 1607). On the other hand, the adaptation of Färe and Grosskopf's (1985) scale efficiency method to determine global qualitative information on the scale-economies regime raises several difficulties in a non-convex setting. Not only it is unable to provide a local measure of the degree of scale economies, but it also cannot account for the possible occurrence of global sub-constant scale economies, i.e., the case where the same level of the constant returns-to-scale cost can be achieved by both increasing and decreasing the current scale size. This latter phenomenon has been introduced and theoretically discussed by Podinovski (2004) only in the context of returns to scale, with Cesaroni *et al* (2017) offering a first empirical exploration of the framework in question, while application to cost analysis still remains to be explored.

Besides the theoretical interest in filling the above gaps for non-convex technologies, we believe that there is a more compelling and practical interest in proceeding this way. In fact, as remarked by Tone and Sahoo (2003), "most of the real-life production processes fail to satisfy these stringent" convexity criteria, on account of several reasons such as overheads, indivisibilities in capital equipment, process

*Correspondence: Giovanni Cesaroni, Department for Economic Policy, Prime Minister's Office, Via della Mercede 9, 00187 Rome, Italy.
E-mail: g.cesaroni@governo.it

[1]This expression, along with "cost-based returns to scale", is due to Sueyoshi (1999, p. 1593). In the rest of the article we use returns to scale and scale economies to denote production-based and cost-based returns to scale, respectively.

indivisibilities due to a different task length associated with each single stage of the production process. Grifell-Tatjé and Kerstens (2008) provide some evidence on the non-convex nature of electricity distribution, while Hackman (2008, pp. 125–133) describes explicit examples of non-convex technologies that arise in resource allocation, producer budgeting and Data Envelopment Analysis with lower bounds.

The purpose of this work is threefold. First, we discuss the analytical problems created by the implementation of Färe and Grosskopf's (1985) approach in a non-convex technology. Second, following the approach of Banker and Thrall (1992), we introduce a new convenient method for the estimation of scale economies which relies on the minimization of the ray average cost of an output mix, which overcomes the difficulties of the previous approach. Third, we illustrate the application of the proposed classification procedure on a data set with the aims of checking for the presence of global sub-constant scale economies and of determining the behavior of local scale economies. In addition, this empirical illustration permits to test whether or not the divergence between global and local indicators—pointed out by Podinovski (2004) in a production setting—extends to cost analysis.

This paper is structured as follows. Section 2 introduces the non-convex technology and examines the difficulties of the Färe and Grosskopf (1985) method. Section 3 presents our new method and briefly discusses its specific features in relation to the problems at hand. Section 4 illustrates the empirical application to a representative sample of Italian local public transit companies. Section 5 offers some conclusions and raises some issues for future research.

2. FDH and the Färe and Grosskopf (1985) approach to scale economies

2.1. Preliminary definitions

The production possibility set we consider is the FDH of the observed production possibilities. Introducing notation, we have n observations, indexed by j ($j = 1, \ldots, n$), using m inputs $x_{ij} (i = 1, \ldots, m)$ to produce s outputs, $y_{rj} (r = 1, \ldots, s)$. The observed input and output vectors are $\mathbf{x}_j = (x_{1j}, \ldots, x_{mj})' \geq \mathbf{0}$ and $\mathbf{y}_j = (y_{1j}, \ldots, y_{sj})' \geq \mathbf{0}$, respectively, with $\mathbf{x}_j, \mathbf{y}_j \neq \mathbf{0}$, and where the prime indicates the transposition operation. Denoting the $m \times n$ matrix of inputs as $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_n]$ and the $s \times n$ matrix of outputs as $\mathbf{Y} = [\mathbf{y}_1, \ldots, \mathbf{y}_n]$, then the general form of the production possibility set can be expressed as

$$T_K = \left\{ (\mathbf{x}, \mathbf{y}) \,|\, \mathbf{Xz} \leq \mathbf{x}, \mathbf{Yz} \geq \mathbf{y}, z_j \equiv w\lambda_j, \lambda_j \in \{0, 1\}, \sum_{j=1}^{n} \lambda_j = 1 \right\} \tag{1}$$

where $\mathbf{z}$ is the $n \times 1$ vector with components equal to $w\lambda_j$, and $w > 0$ is a scaling factor which introduces different returns to scale assumptions in the technology, indexed by K. These are the following: $w > 0 \Leftrightarrow T_{\text{CRS}}$, $w = 1 \Leftrightarrow T_{\text{VRS}}$, $w \leq 1 \Leftrightarrow T_{\text{NIRS}}$, $w \geq 1 \Leftrightarrow T_{\text{NDRS}}$, where the abbreviations CRS, VRS, NIRS and NDRS stand for constant, variable, non-increasing and non-decreasing returns to scale, respectively. Note that T_{VRS} is the least restrictive reference technology among the listed ones, in that it generates the smallest set enveloping the original observations.

As far as costs are concerned, we denote the vector of input prices as $\mathbf{p}_j = (p_1, \ldots, p_m) > \mathbf{0}$ with $\mathbf{p}_j\mathbf{x}_j$ representing the total cost of observation j for producing its output vector $\mathbf{y}_j$, which also represents its current scale size.[2] In any of the reference technologies listed above, the cost efficiency of this unit can be evaluated by solving the following mixed-integer programming problem:

$$\begin{aligned} &\underset{(\mathbf{x}_h, \mathbf{y}_h, z_h)}{\text{Min}} \ \frac{\mathbf{p}_j\mathbf{x}_h z_h}{\mathbf{p}_j\mathbf{x}_j} \\ &\text{s.t.} \\ &\mathbf{y}_h z_h \geq \mathbf{y}_j \end{aligned} \tag{2}$$

where $h = 1, \ldots, n$ denotes a generic observation. A solution to program (2), $(\mathbf{x}^*, \mathbf{y}^*, z^*)$, determines the cost efficiency score of observation j as $\frac{\mathbf{p}_j\mathbf{x}^* z^*}{\mathbf{p}_j\mathbf{x}_j}$, which is equal to 1 for a cost-efficient DMU.

We conclude this subsection with an important definition.

Definition 1 For an observation j global sub-constant scale economies occur when in T_{CRS}, $\frac{\mathbf{p}_j\mathbf{x}^* z^*}{\mathbf{p}_j\mathbf{x}_j} < 1$ and there are at least two solutions z_1^* and z_2^* such that $\mathbf{p}_j\mathbf{x}_1^* z_1^* = \mathbf{p}_j\mathbf{x}_2^* z_2^* = \mathbf{p}_j\mathbf{x}^* z^*$ with $z_1^* < 1$ and $z_2^* > 1$.

In other words, global sub-constant scale economies are the case where a scale inefficient unit has the same CRS optimal cost determined by two different scale sizes, one which is larger and the other which is smaller than the current scale size: $\mathbf{y}_1^*$ and $\mathbf{y}_2^*$, respectively, where $\mathbf{y}_1^* z_1^* \geq \mathbf{y}_j$ and $\mathbf{y}_2^* z_2^* \geq \mathbf{y}_j$.

2.2. The Färe and Grosskopf (1985) approach

The method developed in Färe and Grosskopf (1985) relies on the computation of cost-scale efficiency—defined as the ratio between CRS and VRS cost efficiency scores—of the points lying on the frontier of the technology. If the examined point is cost-scale efficient, then it exhibits CRS. Otherwise, a third NIRS reference technology is used to establish qualitative information on the nature of scale inefficiency, i.e., on the direction to the VRS optimal scale size whose projection determines the CRS cost efficiency score. In particular, if

[2] A scale size variable indicates the level at which either inputs or outputs are actually being employed by a unit under evaluation (i.e., current). In the analysis of scale economies, this scale size variable is normally expressed in terms of the outputs (see, e.g., Färe and Grosskopf, 1985, p. 600). We follow this convention (as moreover discussed in Section 3).

NIRS and VRS cost efficiencies are equal, then scale inefficiency is due to DRS (the optimal scale size is lower than the current one), otherwise it is due to IRS (the optimal scale size is greater than the current one). Podinovski (2004) denotes the resulting classification as global because it is based on the absolute minimum cost (i.e., the CRS cost) and is determined by a scale size which may be rather distant from the current under examination. As such, the method does not provide quantitative information relating to the degree of scale economies, which is commonly measured as a scale elasticity—the ratio of marginal to average cost.

The straight application of this method to a non-convex technology may encounter the same problem documented by Podinovski (2004, p. 242) in the analysis of returns to scale to production. Essentially, the occurrence of multiple optima in the CRS cost efficiency problem can lead to a wrong classification of the global scale-economies regime of some observations: These could be classified as enjoying increasing scale economies, while they actually operate under a sub-constant regime.

A simple example can be used to illustrate both the meaning of global sub-constancy in cost analysis and the classification error in question. In a three-dimensional space, consider three observations having a two input–one output vector (x_{1j}, x_{2j}, y_j) and an identical input price vector (p_1, p_2): DMU A $(1, 1.5, 1.5)$, DMU B $(2, 3, 3)$, DMU C $(2, 1, 2)$ with input price vector $(2, 1)$. Clearly, A and B are proportional replicas of each other. The projection of the VRS Free Disposal Hull T_{VRS} on the (x_1, x_2) plane and its three-dimensional representation are illustrated in Figures 1 and 2, respectively.

From Figures 1 and 2, it can be easily seen that each of the three fictitious observations is both a VRS cost-efficient unit and a most productive scale size (i.e., CRS technically efficient). Nevertheless, while A and B operate under global constant scale economies, because their CRS cost efficiency score is equal to 1, this is not true for C.[3] In fact, C can decrease the ray average cost of its production $\left(5/2\right)$ to the optimal level equal to $7/3$, by adopting both the input mix and the output scale size of either A or B. In other words, its CRS optimal cost is lower than its VRS cost: $7/3 \cdot 2 < 5$. This means that C operates under global sub-constant scale economies in that it can either decrease or increase the scale of production to reach the minimum ray average cost along ray OB (see Figure 2).

This numerical example clarifies that the problem of classifying global scale economies is in principle different and more complex than the one of returns to scale. The presence of allocative inefficiency in the input mix makes the achievement of maximal ray average productivity, the CRS technical-efficiency condition, not sufficient to obtain global

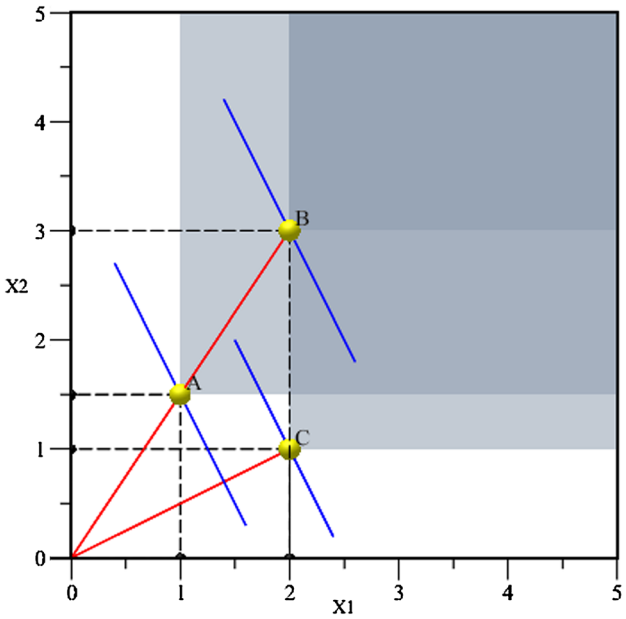

Figure 1 Numerical example of cost efficiency in a two-input section.

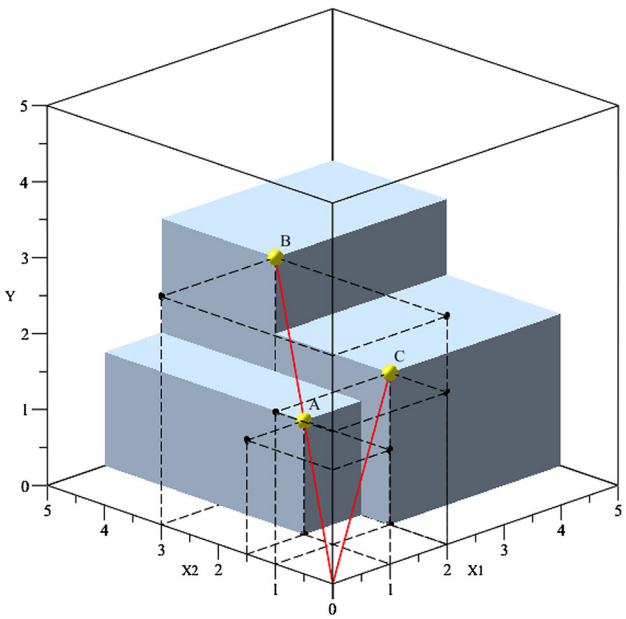

Figure 2 Numerical example of cost efficiency in a two input-one output space.

constant scale economies. This is paradigmatically shown by the case of DMU C, which is CRS technically efficient and whose input mix differs from those of DMUs A and B (for a more profound theoretical discussion, see Cesaroni and Giovannola, 2015, Section 3).

Turning back to the classification based on Färe and Grosskopf's (1985) method, we can remark that DMU C has a NIRS cost efficiency score equal to $14/15$, which differs from its VRS cost efficiency score that equals unity. According to

[3]The calculation of CRS and NIRS scores according to model (2) is shown in Appendix 1. The following discussion exploits the equivalence mentioned later on at point 2. Section 3.1 to illustrate the CRS cost efficiency score in terms of ray average cost.

the above-told method, DMU C exhibits global increasing scale economies, while it actually is characterized by global sub-constant scale economies.

The difficulties raised by the presence of multiple optimal solutions can only partially be solved by bringing the NDRS technology into the picture of scale economies determination. This is in the same spirit as Kerstens and Vanden Eeckaut's (1999) method which employs three reference technologies: NIRS, NDRS and CRS. The error in the global classification is still present, because unit C would be classified as constant instead of the sub-constant characterization. Fortunately, this shortcoming can be corrected by introducing the same qualifications suggested by Podinovski (2004) for the returns to scale case. These qualifications consist in the use of the VRS (instead of the CRS) technology, and in the explicit consideration of the sub-constant case. In this setting, global sub-constant scale economies occur if and only if NIRS and NDRS cost efficiency scores are equal and strictly lower than the VRS score. Note that in our example DMU C satisfies exactly these conditions.

We can conclude that the amended Kerstens and Vanden Eeckaut's (1999) method for the measurement of global scale economies is computationally complex, because it requires the solution of three mixed-integer programming problems (NIRS, NDRS, VRS). Moreover, just like Färe and Grosskopf's (1985) procedure, it cannot provide quantitative information on the degree of scale economies in a neighborhood of a DMU's scale of operations. The next section defines a method which provides a solution to all these difficulties discussed so far.

3. A ray average cost approach

Cesaroni and Giovannola (2015) show that a dual measure of returns to scale can be based on the minimization of the ray average cost on a VRS technology, both convex and non-convex, satisfying the assumption of strong disposability only. These authors in fact extend Banker's (1984) and Banker and Thrall's (1992) analyses by allowing for a variable input mix and a non-convex technology. In the following, we propose a method for FDH technologies with special attention to the inclusion of multiple solutions in the optimization program—which induces the sub-constant case—and to its analogy with the above-mentioned analyses. The general concept of scale economies we employ does not deviate from the standard approach, which ascertains the behavior of the cost function in response to a variation of outputs at given input prices[4] (see, e.g., Baumol *et al*, 1982). Herein, constant scale economies are said to occur when the average cost is stationary at a minimum level, otherwise—given the convexity of the technology—we have increasing or decreasing scale economies when the average cost is decreasing with an increase or a decrease in the output's scale size, respectively (see Section 4.2.1 in Sueyoshi, 1999, and Section 4.2 in Tone and Sahoo, 2003). In this approach, scale economies are a result of the technological and organizational factors which define/determine the frontier of the production possibility set at different levels of output (see Silberton's definition in Tone and Sahoo, 2003, p. 167), as an example: an higher quantity of input utilization, due to larger plants and associated higher variable-factors, may bring about a lower average cost at an output's scale size which is larger than the current one.

[4]In principle our method allows for a variation of input prices associated to a change in the scale size (e.g. bulk buying of inputs may lower their prices) but this choice implies that an optimal scale size is not necessarily a most productive scale size, i.e. it is not CRS technically efficient. For a discussion see Cesaroni and Giovannola (2015, Sect. 4.3).

3.1. A new classification method for scale economies

Evaluation of the overall and scale efficiency of the output mix of a DMU j can be accomplished by means of a ray average cost ratio or average-cost efficiency (ACE): The ratio between the ray average cost evaluated at a unit $(\mathbf{x}_h, \mathbf{y}_h) \in T_{VRS}$ and that evaluated at the DMU's current scale size. Following Cesaroni and Giovannola (2015, pp. 122–123), this ACE ratio can be expressed as follows

$$R_j \equiv \frac{\mathbf{p}_j \mathbf{x}_h}{\mathbf{p}_j \mathbf{x}_j} \cdot \gamma_{j,h} \tag{3}$$

where $\gamma_{j,h}$ is the radial scaling factor obtained from the comparison between the output vector of j and that of h. This radial scaling factor is computed as follows

$$\gamma_{j,h} = \max_r \left\{ \frac{y_{rj}}{y_{rh}} \right\}, \text{ where } \gamma_{j,h} \in (0, \infty] \tag{4}$$

For an exogenously chosen unit h, the ACE ratio R_j indicates the gain in average cost that a single DMU j achieves if it changed its scale size by adopting $\mathbf{x}_h$ to produce $\mathbf{y}_h$—or equivalently the radial projection of its output vector onto the scale size of unit h, $\frac{1}{\gamma_{j,h}} \cdot \mathbf{y}_j$. Note that $\mathbf{x}_h$ is arbitrary and not a radial projection of the input mix $\mathbf{x}_j$, since input proportions are allowed to vary. Moreover, remark that $\frac{1}{\gamma_{j,h}} \cdot \mathbf{y}_j \leq \mathbf{y}_h$ such that the VRS assumption is not violated.

The efficiency score of a given DMU j is obtained from the minimization of (3) over T_{VRS}, determining the average-cost efficiency measure (ACE) R_j^*. More importantly, it is proven that the minimizer of this optimization program, which we call an optimal scale size (OSS) o, has the following important properties (see Cesaroni and Giovannola, 2015, Propositions 3 and 7 resp.):

(1) An OSS is average-cost efficient: $R_o^* = 1$;
(2) The ACE measure R_j^* is equal to CRS cost efficiency and can be decomposed into the product of VRS cost efficiency and cost-scale efficiency.

In other words, our optimal scale size minimizing the ray average cost under VRS coincides with the scale size that

minimizes total cost of production under the CRS assumption, i.e., R_j^* is equal to $\frac{\mathbf{p}_j\mathbf{x}^*z^*}{\mathbf{p}_j\mathbf{x}_j}$ obtained in T_{CRS} (see Definition 1). This implies that the ACE measure can in principle be used to estimate the global scale economies of a generic DMU. To this purpose, we only need to consider the information given by coefficients R_j^* and $\gamma_{j,o}$. The method for the classification of scale economies of cost-efficient points in T can be formulated as follows:

Proposition 1 *For a cost-efficient DMU j, we have*

(i) $R_j^* = 1$, *and* $\gamma_{j,o} = 1$ *in a solution, then global constant scale economies prevail*
(ii) $R_j^* < 1$ *and* $\gamma_{j,o} < 1$, *then global increasing scale economies prevail*
(iii) $R_j^* < 1$ *and* $\gamma_{j,o} > 1$, *then global decreasing scale economies prevail*
(iv) $R_j^* < 1$ *and both* $\gamma_{j,o} > 1$ *and* $\gamma_{j,o'} < 1$ *in any pair of different solutions, then global sub-constant scale economies prevail.*

Proof The possible existence of multiple solutions to the minimization of (3) implies that multiple values of $\gamma_{j,o}$ might be associated to a given R_j^*. Therefore, (i), which includes the case[5] of $R_j^* = 1$ and $\gamma_{j,o} \neq 1$, can be immediately derived from the above properties under (1) and (2), while (ii), (iii) and (iv) follow from: expression (4), the impossibility of the simultaneous occurrence of $R_j^* < 1$ and $\gamma_{j,o} = 1$ (see Cesaroni and Giovannola, 2015, Appendix A.9). This ends the proof.

Based on the properties under point (2) above, two important characteristics of the method can be pointed out. First, j being a cost-efficient point in T, R_j^* represents its cost-scale efficiency: in this sense, our method is based on the measurement of cost-scale efficiency (just like the one of Färe and Grosskopf, 1985). Second, the equivalence of ACE and CRS cost efficiency programs implies that $\gamma_{j,o} = z_j^*$, where the second member denotes the CRS solution to program (2) and is equivalent to $\sum_{j=1}^{n}\lambda_j^*$, the sum of weights in the optimal solution in Banker and Thrall's (1992, p. 81) method.

As far as the sub-constancy case is concerned, expressions (3) and (4) reveal that exact proportionality of two different OSSs is a sufficient condition for obtaining it at a unit j which is not cost-scale efficient. Accordingly, if exactly proportional replicas of the observations acting as an OSS are present in the data set, or rather are assumed to exist on the basis of an *elementary replicability* postulate (see Tulkens, 1993, p. 191; Agrell and Tind, 2001, p. 132), then we can empirically establish the occurrence of the sub-constancy case.

[5]See Case 2 in Banker and Thrall (1992), p. 81. A numerical illustration is given in Appendix 2.

3.2. Discussion

Measure R_j represents—by construction—the ratio between average costs of two different VRS scales of production. Therefore, it immediately supplies quantitative information on the degree of scale economies, which is unavailable in the Färe and Grosskopf (1985) approach. Moreover, the use of a ratio of two average costs avoids the need to compute a scale elasticity measure. A ratio of average costs provides at the same time information whose meaning is unambiguous and more useful from the managerial point of view compared to the practical indeterminacy of the concept of marginal cost in a non-convex setting.

In our approach, $1 - R_j^*$ indicates the maximum decrease in average cost that can be associated with a discrete variation $1 - \gamma_{j,o}$ in the scale of production. These two magnitudes suffice to completely characterize the global economies of scale of any cost-efficient DMU. This indicates the direction in which the absolute minimum of the ray average cost can be found.

Furthermore, the two ratios in question can also be employed to infer the local behavior of the ray average cost in any neighborhood of this DMU's output vector, by enumerating all of the frontier points which determine a positive degree of scale economies, $1 - R_j > 0$, i.e., points that are relative minima. In this sense, the method is able to supply relevant information about organizational and technical changes which ensure reductions in the average cost in any interval of the current scale of operations.

Also the computational advantages of our approach are noteworthy. It brings to cost analysis the same kind of simplification accomplished by Soleimani-damaneh *et al* (2006) and Soleimani-damaneh and Reshadi (2007) in production analysis. Their methods do not apply to cost analysis and furthermore do not allow for the global sub-constant case, a case which it is wrongly labeled as constant (see Cesaroni *et al*, 2017).

4. Empirical application

We apply the proposed method to a sample of companies operating in the Italian local public transit industry. The sample is especially relevant for two reasons: first, it is the sample used for the estimates regarding the supply-side conditions as presented in the official annual report on the sector[6]; second, it is taken from an industry which is about to undergo a significant restructuring in the scale of activity of individual firms, as a result of a higher degree of market contestability which should follow the recent establishment of the National Transport Authority.

The data set consists of observations taken from the balance sheets of 43 companies in the year 2012, the latest available. These companies account for the 60% of the sector's aggregate

[6]See part 2, p. 41 and following in Isfort *et al* (2014).

supply. Scale economies in transport industries require to take account of the simultaneous expansion of an output, in our case the vehicle-kilometers travelled, and the size of the transport network. As a proxy of this latter network variable, we introduce the area served by a company as a second output. On the input side, we consider three inputs expressed in real terms: the number of company staff, the number of vehicles and the quantity of a composite commodity representing the consumption of fuel, energy, materials and spare parts. Each company faces a specific input price vector made up by: the average wage of its staff, the average depreciation of its stock of vehicles and a price index for the composite intermediate commodity. For confidentiality reasons, names of individual operators cannot be disclosed: When refereeing to any single company, we use the name Decision Making Unit (DMU) and its number in our database.

While a few studies have appeared on the cost efficiency of Italian urban transit, most recent studies use parametric specifications and are therefore little useful for comparative purposes (e.g., Ottoz *et al*, 2009; Piacenza, 2006 among others). The only nonparametric study we are aware of is meanwhile quite dated (see Levaggi, 1994). Therefore, a comparative analysis is hard to make.

4.1. Estimates of global scale economies

This subsection illustrates the results obtained regarding the measurement of global scale economies of individual companies. We first present estimates obtained in the FDH model and then those due to an extension of the model based on a selective replicability postulate, which can be inferred from the former estimates.

In Table 1, we compare cost efficiency and average-cost efficiency scores (CE and ACE, respectively) of the original observations. We can note a remarkable difference between the summary statistics of the two kind of efficiency measures, which denotes a widespread diffusion of cost-scale inefficiency as witnessed by the ratio between the two means reaching a 0.72 value and by the number of efficient units: Only 3 of the 22 cost-efficient units exhibit cost-scale efficiency.

The qualitative information on the nature of scale inefficiency and its breakdown according to the scale size of the main output is shown in Table 2, where the abbreviations DSE, ISE, CSE and GSCE stand for decreasing, increasing, constant, and sub-constant scale economies, respectively. No single case of global sub-constancy has been found, while the 40 cost-scale inefficient units are mainly operating under decreasing rather than increasing scale economies, i.e., their current scale of operations is larger/lower, respectively, than that which minimizes the ray average cost.

Table 3 presents some characteristics of the observations operating under constant scale economies. Two out of the three OSSs have a size lower than 10 million vehicle-km and act as a benchmark for some dominated unit, contrary to the largest OSS (DMU 2) which does not dominate any unit. Moreover, note that DMU 32 is the benchmark for the large majority of cost-scale inefficient units (38 out of 40 times).

Table 1 Comparison of efficiencies

	CE	*ACE*		*CE*	*ACE*
Summary statistics of efficiency scores			*DMU type*		
Minimum	0.4927	0.3278	Inefficient	21	40
Mean	0.8768	0.6321	Efficient	22	3
Maximum	1	1			
Stand. Dev.	0.1609	0.1686			

Table 2 Global scale economies

Vehicle-Km (Millions)	*DSE*	*ISE*	*CSE*	*GSCSE*	*Total*
		(# of units)			
<5	1	7	1	0	9
5–10	1	10	1	0	12
10–20	10	0	0	0	10
20–40	8	0	1	0	9
>40	3	0	0	0	3
Total	23	17	3	0	43

Table 3 Optimal Scale Size (OSS) Characteristics

	Vehicle-Km (millions)	*Service area (km^2)*	*Dominated units (#)*
DMU 9	5.36	2779.34	2
DMU 32	9.2	2275.42	38
DMU 2	31.6	11687.06	0

Moreover, the information displayed in Tables 2 and 3 can be used to point out that the number of firms in a long-run equilibrium of the industry, characterized by CRS cost efficiency at the current aggregate output, is likely to be substantially greater than the existing number of firms. In fact, not only cost-scale inefficient units are concentrated mainly in the decreasing regime, i.e., the scale-size range above that of DMU 32 (9.2 million vehicle-km), but even the average output size of these decreasing regime units exceeds by far that of the units in the increasing regime, operating below DMU 32. In other words, if the aggregate output of the cost-scale inefficient units was to be produced according to the optimal scale size, in order to minimize its total cost, the number of firms should be increased.

Another interesting implication, which is more relevant at the company level, regards the comparison between the cost structure of the observed DMUs (defined by the percentages of the actual cost due to each input, $\frac{\mathbf{p}_i\mathbf{x}_i}{\sum_i \mathbf{p}_i\mathbf{x}_i}$) and that of the prevailing optimal scale size (DMU 32). The results are shown in Table 4

Table 4 Actual and optimal cost structures

	Input 1	*Input 2*	*Input 3*
Average difference	0.116	−0.046	−0.070
Number of positives	42	4	1
Number of negatives	0	38	41

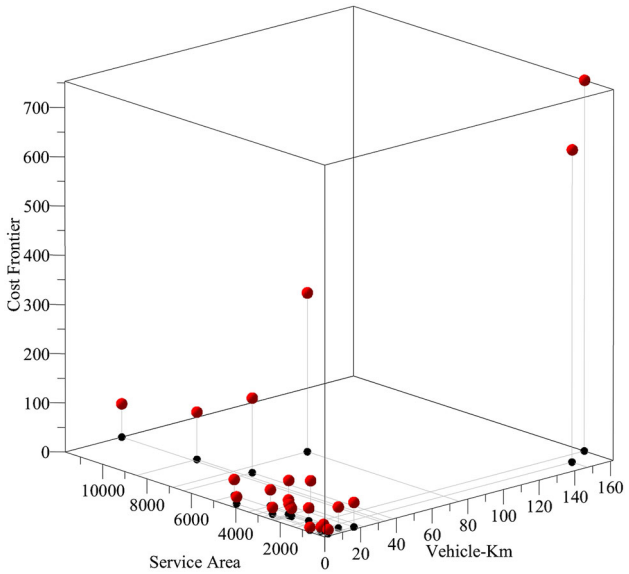

Figure 3 Estimated cost function.

The difference between the actual and the optimal cost structure is very significant both in its uniformity (number of positives/negatives) and in its amount. We remark that, when compared to the prevailing optimal scale size, nearly all DMUs exhibit: a higher percentage of staff cost (+11.6 percentage points on average), a lower percentage on vehicles (-4.6 p.p.) and on the intermediate consumption commodity (-7 p.p.). Therefore, the managerial implications derived from DMU 32 include, in addition to the variation of the output's scale size, the necessity of reorganizing the cost structure by reducing the weight of staff and by increasing that of the number of vehicles (input 2 with a direct impact on input 3 as far as fuel and lubricating oil consumption are concerned) and of their ordinary maintenance (spare parts which are part of input 3).

Quite interestingly, by considering the DMUs projections on the efficient frontier,[7] the following three-dimensional graph illustrates the complex behavior of the estimated cost function in our non-convex technology. In Figure 3, the vertical axis represents the frontier cost, while the axes in the horizontal plane denote the two outputs service area and vehicle-km produced, respectively. The chosen perspective should give an idea of the alternation of concave and convex behavior of the total cost with respect to outputs.

Table 5 Global Sub-Constant Scale Economies (GSCSE) under replicability

Unit	*Lambda 1*	*Lambda 2*	*ACE*
DMU 3	1.1130	0.5565	0.7594
DMU 8	1.3674	0.6836	0.5858
DMU 13	1.9968	0.9984	0.7089
DMU 21	1.0290	0.5145	0.7537
DMU 27	1.9968	0.9984	0.7350
DMU 28	1.5151	0.7575	0.7805
DMU 31	1.1033	0.5516	0.6999
DMU 37	1.1130	0.5565	0.7150
DMU 41	1.4238	0.7119	0.5971
DMU 43	1.9968	0.9984	0.7191

As far as the assumption of replicability is concerned, we believe that it may turn out to be a deceiving hypothesis. In the absence of specific and sound knowledge of the technical and organizational characteristics of the production technology, it can arbitrarily enlarge the number of CRS points of a production technology. Therefore, to limit this kind of risk, we draw on some information on the OSS characteristics and impose the replicability assumption only for those of these units falling in a suitable output range.[8] From Table 3, we conjecture that a relatively safe choice could be to introduce an integer replicability of order 2 only for DMUs 9 and 32. Such an extension of the FDH model yields 10 observations operating under global sub-constant scale economies: These are depicted in Table 5.

The Lambda 1 column reports the $\gamma_{j,o}$ coefficient estimated in the FDH model, while the next column (Lambda 2) presents the analogous coefficient ensuing from the introduction of replicability. The interpretation of these results is rather straightforward. Take as an example DMU 41. While adopting the optimal input mix of its OSS, this unit could either decrease its scale of operations by 43% (measured in terms of its OSS output) or increase it by nearly 29% to obtain a 40% reduction in its average cost.

4.2. Estimates of local scale economies

Integer replicability of an appropriate order can introduce the sub-constant scale economies case, but anyway its validity is questionable. In this subsection, we are showing that global sub-constant scale economies are in practice not necessary to reach the conclusion that a firm, while changing its input mix, can reduce its ray average cost by either increasing or decreasing its scale of operations. To this purpose recall that, for a single cost-efficient DMU j, we can describe the local

[7]We are considering radial-efficient projections of the original observations in the sense of Podinovski (2004, p. 244): see his Definition 5.

[8]On the suitability of an upper bound in the replicability assumption, see Tone and Sahoo (2003, pp. 171–172). Mairesse and Vanden Eeckaut (2002) develop a similar argument in an FDH context.

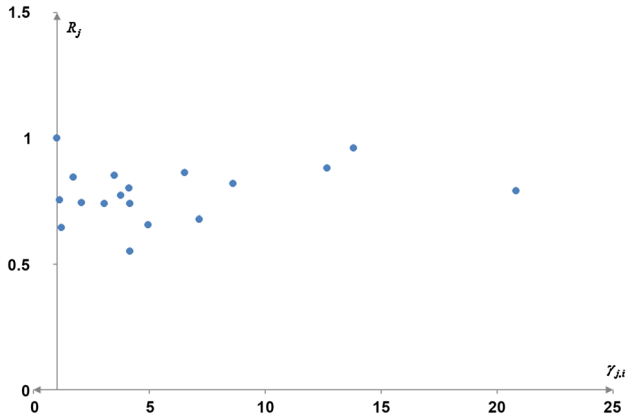

Figure 4 Local scale economies for DMU 5.

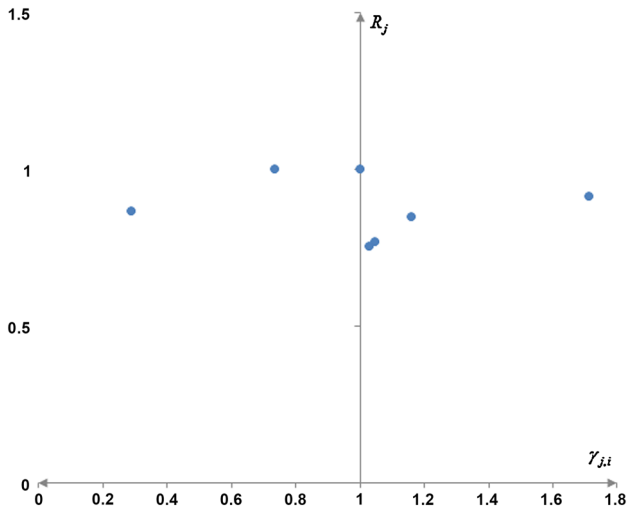

Figure 5 Local scale economies for DMU 21.

behavior of its ray average cost by first computing $(R_j,\gamma_{j,i})$ with respect to each point i belonging to the efficient frontier and then selecting those points that determine a positive degree of scale economies ($R_j < 1$). Such computations have been performed for various units of the sample and reveal that, in a non-convex technology, the global and local indicators of scale economies can be in contrast with each other. In general terms, we can define this contrast as a situation in which there exists at least a relative minimum in the ray average cost which is located in the opposite direction with respect to that of the absolute minimum.

In the following figures, we illustrate the different types of behavior of the ray average cost in the original FDH model for various units. Each DMU under examination is located at the point having coordinates (1,1), with R_j and $\gamma_{j,i}$ being represented on the vertical and the horizontal axis, respectively.

Figure 4 reports a situation for DMU 5 in which the contrast between local and global indicators does not occur. In fact, it

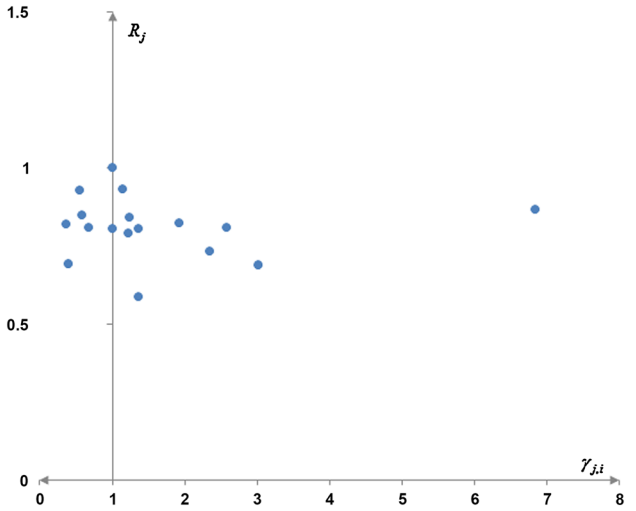

Figure 6 Local scale economies for DMU 8.

can be observed that each of the frontier points delivering a ray average cost lower than one has $\gamma_{j,i} > 1$: Both the global and the local indicators point to the decreasing economies of scale direction.

Conversely, Figure 5 illustrates some moderate contrast for DMU 21. The global indicator is clearly situated in the decreasing direction, slightly to the right of 1; nevertheless, there exist two points in the increasing direction: The farthest on the left gives a ray average cost (0.867) which is halfway between 1 and the absolute minimum (0.754).

The case of a deep contrast is well described in Figure 6 that illustrates DMU 8. While the global indicator is in the decreasing direction (between 1 and 2), we can note that: a) there are five points in the increasing direction; b) the most efficient point in the increasing direction delivers a ray average cost of 0.694, which is rather close to the absolute minimum (0.586).

Finally, Figure 7 illustrates for DMU 12 a case of contrast between globally increasing and locally decreasing economies of scale. The absolute minimum of the ray average cost is clearly in the increasing direction ($\gamma_{j,i} < 1$) and is approximately equal to 0.5, but there is a point in the opposite direction which is the nearest relative minimum and ensures a 0.72 ray average cost.

Besides the general possibility of obtaining reductions in the ray average cost by both increasing and decreasing the scale of operations, the comparison between local and global indicators of scale economies illustrated so far brings to light two more specific managerial implications. First, significant reductions in the ray average cost that may not differ much from the one corresponding to the absolute minimum (global indicator) can be achieved within a smaller range of variation of the output's scale size (e.g., see the case of DMU 5). Second, this latter variation can be in the opposite direction with respect to that of the global indicator (e.g., see the case of DMU 12).

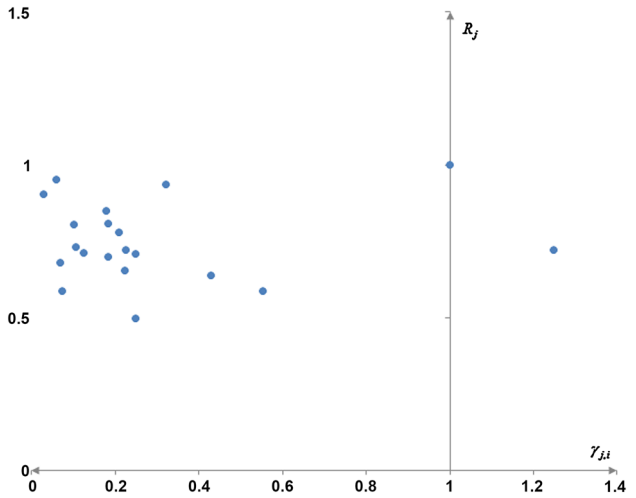

Figure 7 Local scale economies for DMU 12.

Overall, the characteristics pointed out above in this section become especially important when one considers that in the short-run full adjustment of the firm's scale of operations to the optimum determined by the global indicator may be hindered by the presence of adjustment costs, financial and market-demand constraints.

5. Conclusions

This work has introduced a convenient method for the classification of scale economies in non-convex production models, which takes into account the global sub-constant scale economies case and solves the difficulties typical for the approach of Färe and Grosskopf (1985). The application of this method has proven empirically that the contrast between global and local indicators [as revealed by Podinovski (2004)] extends to cost analysis. As far as the managerial implications are concerned, the empirical analysis has, moreover, found out that this kind of contrast can *de facto* provide an individual firm with a wide menu of choices for restructuring its scale of operations and input mix so as to achieve significant reductions in unit costs of production.

Future research will have to investigate the existence of necessary conditions for the occurrence of sub-constant scale economies, as well as the possible different behavior of the ray average cost in convex and non-convex technologies—which may well prove that the sub-constant case and the contrast between local and global indicators cannot occur in convex analysis.

References

Agrell PJ and Tind J (2001). A dual approach to nonconvex frontier models. *Journal of Productivity Analysis* **16**(2):129–147.

Banker RD (1984). Estimating most productive scale size using Data Envelopment Analysis. *European Journal of Operational Research* **17**(1):35–44.

Banker RD, Chang H and Cooper WW (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.

Banker RD, Charnes A and Cooper WW (1984). Some models for estimating technical and scale inefficiencies in Data Envelopment Analysis. *Management Science* **30**(9):1078–1092.

Banker RD and Thrall RM (1992). Estimation of returns to scale using Data Envelopment Analysis. *European Journal of Operational Research* **62**(1):74–84.

Baumol WJ, Panzar JC and Willig RD (1982). *Contestable Markets and the Theory of Industry Structure*. Harcourt Brace Jovanovich: New York.

Cesaroni G and Giovannola D (2015). Average-cost efficiency and optimal scale sizes in non-parametric analysis. *European Journal of Operational Research* **242**(1):121–133.

Cesaroni G, Kerstens K and Van De Woestyne I (2017). Global and local scale characteristics in convex and nonconvex nonparametric technologies: A first empirical exploration. *European Journal of Operational Research*. doi:10.1016/j.ejor.2016.10.030

Deprins D, Simar L and Tulkens H (1984). Measuring labor efficiency in post offices. In M. Marchand, P. Pestieau, H. Tulkens (Eds). *The Performance of Public Enterprises: Concepts and Measurements* (pp. 243–267). North Holland: Amsterdam.

Färe R, Grosskopf S and Lovell CAK (1983). The structure of technical efficiency. *Scandinavian Journal of Economics* **85**:181–190.

Färe R and Grosskopf S (1985). A nonparametric cost approach to scale efficiency. *Scandinavian Journal of Economics* **87**(4):594–604.

Färe R, Grosskopf S and Lovell CAK (1985). *The Measurement of Efficiency of Production*. Kluwer-Nijhoff: Boston.

Grifell-Tatjé E and Kerstens K (2008). Incentive regulation and the role of convexity in benchmarking electricity distribution: Economists versus engineers. *Annals of Public and Cooperative Economics* **79**(2):227–248.

Hackman S (2008). *Production Economics: Integrating the Microeconomic and Engineering Perspectives*. Springer: Berlin.

Isfort, Anav and Asstra (2014). XI Rapporto sulla mobilità in Italia, Roma.

Kerstens K and Vanden Eeckaut P (1999). Estimating returns to scale using non-parametric deterministic technologies: A new method based on the goodness of fit. *European Journal of Operational Research* **113**(1):206–214.

Levaggi R (1994). Parametric and non-parametric approach to efficiency: The case of urban transport in Italy. *Studi Economici* **53**:67–88.

Mairesse F and Vanden Eeckaut P (2002). Museum assessment and FDH technology: Towards a global approach. *Journal of Cultural Economics* **26**(4):261–286.

Ottoz E, Fornengo G and Di Giacomo M (2009). The impact of ownership on the cost of bus service provision: An example from Italy. *Applied Economics* **41**(3):337–349.

Panzar W and Willig RD (1977). Economies of scale in multi-output production. *Quarterly Journal of Economics* **91**(3):481–493.

Piacenza M (2006). Regulatory contracts and cost efficiency: Stochastic frontier evidence from the Italian local public transport. *Journal of Productivity Analysis* **25**(3):257–277.

Podinovski V (2004). Efficiency and global scale characteristics on the "No free lunch" assumption only. *Journal of Productivity Analysis* **22**(3):227–257.

Soleimani-damaneh M, Jahanshahloo GR and Reshadi M (2006). On the estimation of returns to scale in FDH models. *European Journal of Operational Research* **174**(2):1055–1059.

Soleimani-damaneh M and Reshadi M (2007). A polynomial-time algorithm to estimate returns to scale in FDH models. *Computers and Operations Research* **34**(7):2168–2176.

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**(11):1593–1608.

Tone K and Sahoo BK (2003). Scale, indivisibilities and production function in Data Envelopment Analysis. *International Journal of Production Economics* **84**(2):165–192.

Tulkens H (1993). On FDH analysis: Some methodological issues and applications to retail banking, courts and urban transit. *Journal of Productivity Analysis* **4**(1):183–210.

Appendix 1: Derivation of cost efficiency scores of the numerical example

Here we present the derivation from model (2) of the cost efficiency scores shown in Section 2.2. Consider first DMU A, i.e., $j = A$, and compute the cost ratios $\frac{\mathbf{p}_j\mathbf{x}_h}{\mathbf{p}_j\mathbf{x}_j}$ for $h = A, B, C$, obtaining, respectively, $1, \frac{7}{3.5}, \frac{5}{3.5}$. The z_h coefficient can be determined as $\frac{y_j}{y_h}$ which satisfies the output constraint $y_h z_h \geq y_j$ with the equality sign. In the case under examination, we have $z_A = 1$, $z_B = \frac{1.5}{3}$, $z_C = \frac{1.5}{2}$ and $\min \frac{\mathbf{p}_A\mathbf{x}_h z_h}{\mathbf{p}_A\mathbf{x}_A} = 1$ for z_A and z_B: both coefficients are a CRS solution to problem (2) of DMU A because no constraint has been imposed on z_h. The same procedure can be applied to DMU B, i.e., $j = \mathrm{B}$, thus obtaining a CRS cost efficiency score equal to 1, while for DMU C it can be easily checked that this score is $\min \frac{\mathbf{p}_C\mathbf{x}_h z_h}{\mathbf{p}_C\mathbf{x}_C} = \frac{14}{15}$—corresponding to $z_A = \frac{2}{1.5}$ and $z_B = \frac{2}{3}$.

Now we turn to the calculation of the NIRS cost efficiency score of DMU C, $j = \mathrm{C}$. The cost ratios for $h = \mathrm{A,B,C}$ are, respectively, $\frac{3.5}{5}, \frac{7}{5}, 1$. The NIRS technology requires $z_h \leq 1$; therefore, we consider only $z_B = \frac{2}{3}$ and $z_C = 1$ which yields the NIRS cost efficiency score $\min \frac{\mathbf{p}_C\mathbf{x}_h z_h}{\mathbf{p}_C\mathbf{x}_C} = \frac{14}{15}$, corresponding to z_B.

Appendix 2: Illustration of the presence of multiple solutions in case (i)

Condition (i) in Proposition 1 defines the CRS case, which in presence of multiple solutions to the minimization of (3) implies that, in addition to a solution $R_j^* = 1$ and $\gamma_{j,o} = 1$, there may exist some other solution where $R_j^* = 1$ and $\gamma_{j,o} \neq 1$ (see Banker and Thrall, 1992, p. 81). The numerical example of the preceding section—regarding DMU A—illustrates precisely this outcome. In fact, we remark (see Section 3.1) that $\gamma_{j,o} = z_j^*$, where the second member denotes the CRS solution to program (2). As a consequence, DMU A has $R_A^* = 1$ at both $\gamma_{A,A} = 1$ and $\gamma_{A,B} = \frac{1.5}{3}$; according to case (i) it is classified as featuring global constant scale economies.

Received 17 February 2016;
accepted 28 November 2016

Electronic supplementary material The online version of this article (doi:10.1057/s41274-016-0162-7) contains supplementary material, which is available to authorized users.