European Journal of Operational Research 318 (2024) 549–559

Contents lists available at ScienceDirect

European Journal of Operational Research

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

Decision Support

Data envelopment analysis: From non-monotonic to monotonic scale elasticities

Andreas Dellnitz [a], Madjid Tavana [b,c,1,*]

[a] Leibniz-Fachhochschule School of Business, Hannover, Germany
[b] Business Systems and Analytics Department, Distinguished Chair of Business Analytics, La Salle University, Philadelphia, USA
[c] Business Information Systems Department, Faculty of Business Administration and Economics, University of Paderborn, Paderborn, Germany

ARTICLE INFO

Keywords:
Data envelopment analysis
Scale elasticity
Monotonicity
Multiplicative models
Variable returns to scale

ABSTRACT

The concept of returns to scale (RTS) or local scale elasticities in data envelopment analysis (DEA)—stemming from variable returns to scale (VRS) technology—has been recently criticized because of its misbehavior in the case of decreasing returns to scale (DRS). Here, the instrument should imply a downsizing force for improving productivity. In classical VRS technologies, however, it can hide respective improvement potentials: the more, the larger a company is. The non-monotonic behavior of local scale elasticities can address this effect. This study shows this phenomenon does not apply when using multiplicative DEA models. Therefore, we propose a new global scaling index that works in the classical VRS technology. We prove the new index is weakly monotonic and illustrate our theoretical findings in a banking context.

1. Introduction

To remain competitive, for-profit and non-profit organizations must continually enhance their efficiency, improve performance, and show potential productivity gains. Expansion and downsizing are often used as practical strategies to achieve these goals.

Data envelopment analysis (DEA) is a well-established technique for measuring efficiency and productivity among a group of comparable entities—called decision-making units (DMUs)—since the notable contributions of Charnes, Cooper, and Rhodes as well as Banker, Charnes, and Cooper (CCR and BCC), cf. Charnes et al. (1978) and Banker et al. (1984). In recent years, especially the BCC technology—also called variable returns to scale (VRS) technology—has attracted a lot of attention, as sizing decisions can be motivated by this technology inferring scale elasticities of the DMUs under consideration (e.g., Banker et al., 1984; Banker & Thrall, 1992; Førsund, 1996; Golany & Yu, 1997; Fukuyama, 2000; Tone, 2001; Butler & Li, 2005; Podinovski et al., 2009; Zarepisheh et al., 2010; Kleine et al., 2016; Rödder et al., 2017; Ren et al., 2021; Dellnitz et al., 2022; and Taleb et al., 2022). However, Dellnitz and Rödder (2021) recently showed that scale elasticity exhibits non-monotonic behavior in the case of decreasing returns to scale (DRS). Rödder et al. (2022) further show that this non-monotonic behavior is not compatible with the classical concept of productivity. The non-monotonicity of the scale elasticity measure is not a pure weakness but arises from the affine-linear equations of technology implementation. From an economic perspective, this shortcoming can, in the worst case, even lead to large companies being falsely attributed a scale elasticity of maximum productivity. This possible misclassification makes classic scale elasticity an unreliable indicator for BCC technologies—even though this economic evaluation criterion made the BCC idea so scientifically successful 40 years ago, with more than 26,000 citations today.

This study introduces several concepts that do not suffer from this deficiency. First, we show that the scale elasticity behaves weakly monotonically when multiplicative models are used. Second, to enable an analyst to study sizing potentials in classical BCC technology, we propose a new global scaling index and prove that it also behaves weakly monotonic—even if an affine-linear face generates the surface of a technology. The new global scaling index measures the distance to the output level of the DMU exhibiting the most productive scale size (MPSS) (Banker, 1984; Podinovski, 2004a, 2004b; Cook & Zhu, 2011; Esfandiar et al., 2022); which is becoming an increasingly important theoretical concept in practical applications of DEA (e.g., Ray, 2007; Kounetas et al., 2009; Hung et al., 2010; Asmild et al., 2013; Lee, 2015;

* Corresponding author at: Business Systems and Analytics Department, Distinguished Chair of Business Analytics, La Salle University, Philadelphia, PA 19141, United States.
E-mail address: tavana@lasalle.edu (M. Tavana).
[1] Web: http://tavana.us/

https://doi.org/10.1016/j.ejor.2024.05.018
Received 8 July 2023; Accepted 7 May 2024
Available online 15 May 2024

Ray, 2015; Assani et al., 2018). This global scaling index provides management with an answer to the extent to which an activity as a whole needs to be proportionally scaled (reduced or expanded) to achieve an optimal level of productivity—in the DEA ductus, the DMU is then scale efficient. Due to its weakly monotonic behavior, this index could serve as a strategic tool to incentivize scaling decisions and could be complemented by successively reducing inefficiencies along this path.

The remainder of this paper is organized as follows: Section 2 lays out the theoretical and conceptual foundation of our contribution. Section 3 is devoted to DEA preliminaries, developing equations to determine scale elasticities in the context of a classical BCC technology (Section 3.1), and justifying the non-monotonicity problem (Section 3.2). Section 4 exhibits the main results on weakly monotonic scale elasticities in multiplicative technologies. Section 5 shows the new global index and proves its monotonicity property. In Section 6, we demonstrate our theoretical results in a banking context and show that the new index addresses the weaknesses of classical scale elasticities. Additionally, we discuss the economic implications of our results in Section 7. Section 8 presents our concluding remarks and future research directions.

2. Theoretical and conceptual foundation

From a theoretical and economic perspective, elasticities have been a standard tool of micro- and macroeconomic theory since Alfred Marshall (1885)—i.e., for more than one hundred years. Here, they are used to motivate decisions on substitutions, expansions, or contractions etc. by organizations or institutions; in other words, they are, among other things, a standard tool for classifying the size of companies, regions, or countries as evidenced by myriads of publications; a quick search on the Web of Science returns over 17,000 publications on elasticities.

Theoretically, scale elasticities point towards organizations or institutions with maximum productivity. In a single input and single output (SISO) production situation, maximum productivity means having the highest output per input, with no overhead overstressing the input. The following two figures show such a situation, where $y \in \mathbb{R}_+$ is the output, and $x \in \mathbb{R}_+$ the input. The first figure illustrates a smooth production function $y = f(x)$, and the second figure is the productivity y/x when wandering along the frontier of the smooth production function.

In the above SISO case, the two figures show maximum productivity y/x when reaching the red dot. As well known, one can find such a point using a ray that starts in the origin; see again Fig. 1.

In classical parametric studies, the above type of function is specified in advance and then estimated by regression to obtain an approximation—but again with a smooth boundary and often with constant productivity gains along the frontier (e.g., Cobb-Douglas production functions). In contrast, DEA is a data-driven approximation of an underlying production function that works based on observations but connects the observations via piecewise linear segments, as illustrated in Fig. 3.

In this approximation based on the points observed, the productivity gain is often estimated by applying the concept of scale elasticity, which is the change rate of the output as a function of the input change rate. Fig. 2 visualizes that the red dot has the highest productivity. To signal this, the scale elasticity is set to one; this means that any deviation would no longer improve the output-input ratio via scaling per scale elasticity—one percent input change yields one percent output change. A problem with piecewise linear approximations now arises when we realize companies "above" maximum productivity, e.g., the blue and green points at the top of Fig. 3. When calculating scale elasticity, both numbers are below one, indicating an improvement potential for reducing both companies: i.e., for each percent input reduction, we lose less than one percent in the output. However, when comparing the two numbers, one realizes that the number for the blue point is closer to one than for the green point. This is counterintuitive because the green point is closer to maximum productivity. Even worse, if we imagine that the blue company becomes bigger and bigger just by extrapolating the path spanned by the green and blue points, its scale elasticity converges to one—but this number should be reserved to indicate the maximum productivity so that the inference process is unique. Consequently, the classical concept of scale elasticity is not working properly in some regions when applying the piecewise linear approximations of DEA; this makes it an unreliable tool—especially in high-dimensional input and output spaces because here we cannot check a company's position visually. These observations suggest two basic attitudes:

a. We can discard the piecewise linear approximation due to its misbehavior and follow the economists by favoring, for example, piecewise Cobb-Douglas production functions as a construction principle (see Section 4).
b. In many cases, however, the first option is not desirable from a practical point of view due to the success of the piecewise linear approximation in DEA-based studies. Eventually, one can try to find a way to avoid the above misjudgment by developing a new index, which can be proven to behave accurately when using piecewise linear approximations (see Section 5). The following figure might indicate the concept of this novel idea:

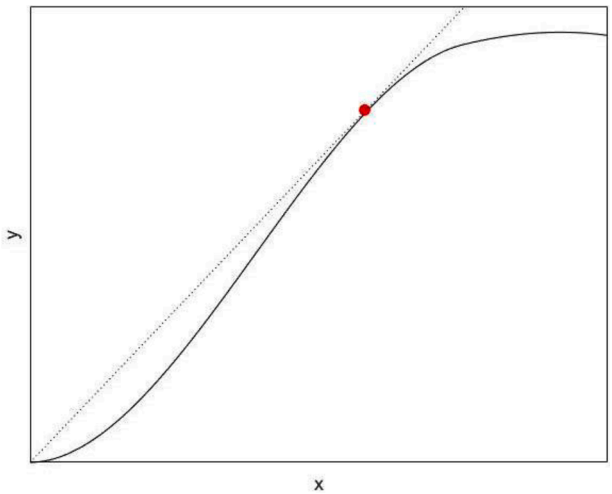

Fig. 1. Smooth production function $y = f(x)$.

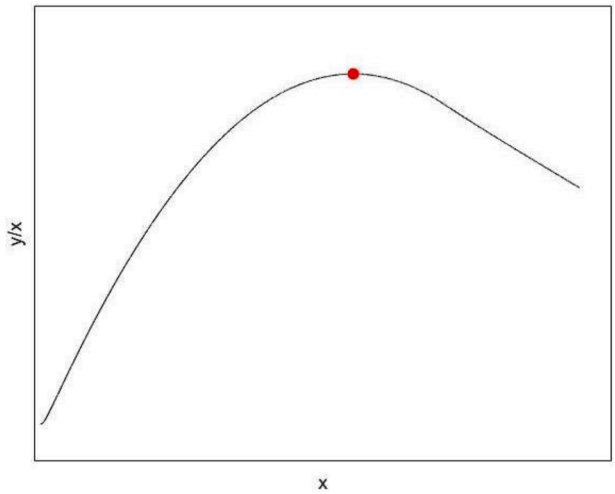

Fig. 2. Productivity along the frontier.

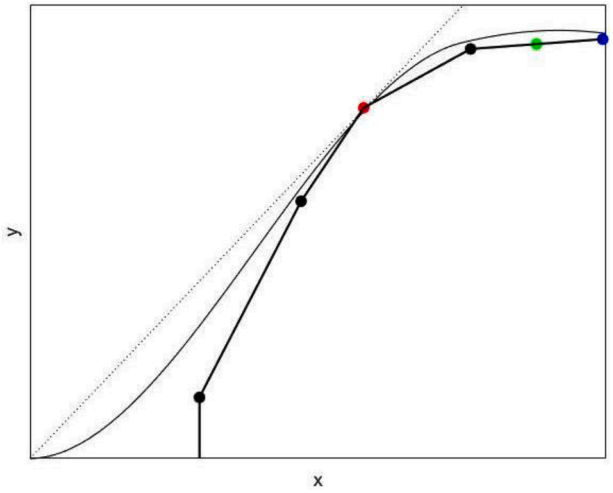

Fig. 3. Approximation of production function $y = f(x)$.

The core concept of our idea is that we calculate the distance to the output level of maximum productivity based on rays crossing the origin, as shown in Fig. 4. This approach is based on the paths determined by the red parts in Fig. 4; one can already guess that the path for the blue company is a little longer than for the green one. In principle, this measure has two significant advantages: First, one can use usual DEA-based calculation results—so we can use established software—to determine the distance. Second, one can prove that this distance behaves monotonically on (high-dimensional) input and output spaces. The latter, in particular, makes the concept a potent tool when examining a context that includes many large companies—as in the banking sector—as here, one is often confronted with the problem that scale elasticities are close to one. Before developing the main results of this paper, we need to revisit the central philosophy of DEA in the next section.

3. Scale elasticities in DEA

3.1. Scale elasticities in BCC models

Scale elasticities as a local index have a long history in DEA since Banker et al. (1984). From a conceptual point of view, this index should point the way towards maximum productivity. However, Dellnitz and Rödder (2021) have shown that this index fails in the most critical cases,

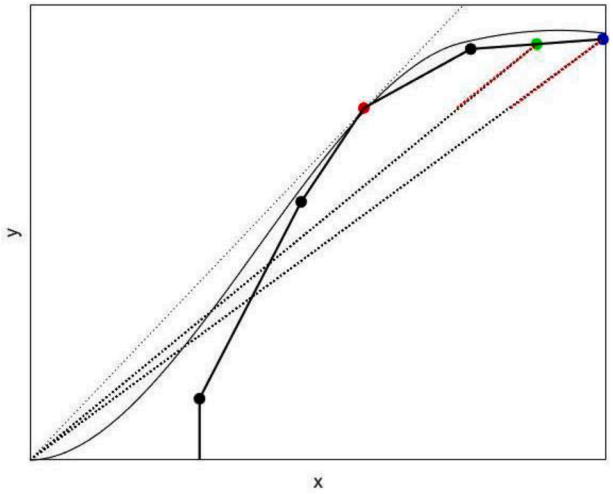

Fig. 4. The concept of the new index.

namely when activities are too large and should shrink because they are very far from maximum productivity. Before pointing out this shortcoming and developing an index that better supports the expansion and downsizing of companies, we will develop important, well-known principles of DEA and scale elasticities in this section.

Although we focus exclusively on the input-oriented efficiency of DMUs so as not to overload the content, the results can easily be applied to other efficiency measures, e.g., output-oriented or slack-based versions. That is, for a DMU $k \in \mathcal{J}$, with $\mathcal{J} = \{1, \ldots, J\}$, we solve the BCC multiplier form (1) to calculate its efficiency; an axiomatic foundation can be found in Banker et al. (1984).

$$\begin{aligned} \max\ g_k = {} & \mathbf{U}_k^{\mathrm{T}}\mathbf{y}_k + u_k \\ & \text{s.t.} \\ & \mathbf{V}_k^{\mathrm{T}}\mathbf{x}_k = 1 \\ & \mathbf{U}_k^{\mathrm{T}}\mathbf{y}_j + u_k - \mathbf{V}_k^{\mathrm{T}}\mathbf{x}_j \le 0 \qquad \forall j \in \mathcal{J} \\ & \mathbf{U}_k, \mathbf{V}_k \ge \mathbf{0} \text{ and } u_k \text{ free} \end{aligned} \tag{1}$$

Let $\left(\mathbf{x}_j,\ \mathbf{y}_j\right) \in \mathbb{R}_+^{M+S} \forall j \in \mathcal{J}$ be the observed activities—with $\mathbf{x}_j = \left(\mathrm{x}_{1j}, \ldots, \mathrm{x}_{mj}, \ldots \mathrm{x}_{Mj}\right)^{\mathrm{T}}$ being the inputs and $\mathbf{y}_j = \left(y_{1j}, \ldots, y_{sj}, \ldots y_{Sj}\right)^{\mathrm{T}}$ the outputs—of all DMUs. $\mathbf{U}_k, \mathbf{V}_k$ are the non-negative vectors of output and input multipliers. Let $\mathbf{U}_k^*, \mathbf{V}_k^*, u_k^*$ be an arbitrary optimal solution concerning (1), then $g_k^* \le 1$ determines the DMU's efficiency and u_k^* indicates its RTS situation: if $u_k^* > 0$ (< 0), we have increasing (decreasing) RTS; a DMU k encompasses constant RTS if $u_k^* = 0$. To justify these statements, one has to prove the following Lemma 1.

Lemma 1. *Let $g_k^*, \mathbf{U}_k^*, \mathbf{V}_k^*, u_k^*$ be an arbitrary optimal solution regarding (1). Then, radial output changes $\mathbf{y}_k \to (1+\varepsilon_k)\mathbf{y}_k$ under infinitesimal radial input changes $\mathbf{x}_k \to (1+\delta)\mathbf{x}_k$ must yield*

$$\frac{\varepsilon_k}{\delta} = \frac{\mathbf{U}_k^{*\mathrm{T}}\mathbf{y}_k + u_k^*}{\mathbf{U}_k^{*\mathrm{T}}\mathbf{y}_k} = 1 + \frac{u_k^*}{\mathbf{U}_k^{*\mathrm{T}}\mathbf{y}_k}, \quad \text{with } \delta \in \mathbb{R}\backslash\{0\}, \tag{2}$$

to maintain the efficiency level g_k^* of DMU k.

Eq. (2) measures the local scale elasticity of DMU k given an arbitrary optimal solution regarding (1). Please note that Eq. (2) is valid only as long as the weights $\mathbf{U}_k^*, \mathbf{V}_k^*, u_k^*$ remain optimal (Kleine et al., 2014). To prove Lemma 1, one has to reorder the following (in-)efficiency equation $\mathbf{U}_k^{*\mathrm{T}}(1+\varepsilon_k)\mathbf{y}_k + u_k^* - g_k^*\mathbf{V}_k^{*\mathrm{T}}(1+\delta)\mathbf{x}_k = 0$ (Dellnitz & Rödder, 2021).

However, the above reasoning may suffer from multiple optimal solutions in Eq. (1). Therefore, Banker and Thrall (1992) have proven that for an efficient ($g_k^* = 1$) DMU k, the optimization of (2) is sufficient to fully fathom a DMU's RTS interval. Dellnitz and Rödder (2021) extended the statement of Banker and Thrall (1992) and have proven that it is also true for an inefficient ($g_k^* < 1$) DMU k.

$$\begin{aligned} u_k^- = \inf u_k \text{ and } u_k^+ = \sup u_k & \\ \text{s.t.} & \\ \mathbf{V}_k^{\mathrm{T}}\mathbf{x}_k = 1 & \\ \mathbf{U}_k^{\mathrm{T}}\mathbf{y}_k + u_k = g_k^* & \\ \mathbf{U}_k^{\mathrm{T}}\mathbf{y}_j + u_k - \mathbf{V}_k^{\mathrm{T}}\mathbf{x}_j \le 0 \qquad \forall j \in \mathcal{J} & \\ \mathbf{U}_k, \mathbf{V}_k \ge \mathbf{0} \text{ and } u_k \text{ free} & \end{aligned} \tag{3}$$

Applying (3) with $\mathbf{U}_k^+, \mathbf{V}_k^+, u_k^+$ and $\mathbf{U}_k^-, \mathbf{V}_k^-, u_k^-$ being respective optimal solutions, we then have:

- $u_k^+ \ge u_k^- > 0 \rightarrow$ increasing RTS (IRS)
- $u_k^- \le u_k^+ < 0 \rightarrow$ decreasing RTS (DRS)
- $u_k^+ \ge 0 \ge u_k^- \rightarrow$ constant RTS (CRS)

Consequently, scale elasticities can vary with $u_k^- \le u_k \le u_k^+$, i.e., we obtain $\frac{\varepsilon_k^-}{\delta} = \frac{\mathbf{U}_k^{-\mathrm{T}}\mathbf{y}_k + u_k^-}{\mathbf{U}_k^{-\mathrm{T}}\mathbf{y}_k}$ and $\frac{\varepsilon_k^+}{\delta} = \frac{\mathbf{U}_k^{+\mathrm{T}}\mathbf{y}_k + u_k^+}{\mathbf{U}_k^{+\mathrm{T}}\mathbf{y}_k}$; this particular effect occurs e.g.

when an efficient DMU's activity—or its projection if $g_k^* < 1$—is on a vertex of the underlying BCC technology or polyhedron.

With the above nomenclature, we can characterize another activity feature: a DMU with MPSS if it is efficient (Banker, 1984) and exhibits CRS (cf. Section 5).

3.2. Non-monotonic scale elasticities in BCC technology

In the latter section, we argued that scale elasticities could vary if an activity or its projection is on a vertex of a BCC technology. However, scale elasticities also vary on the face or edge of such technology due to the measure's dependence on the activity's output level. The next theorem provides this link.

Theorem 1. *Let $g_k^*, \mathbf{U}_k^*, \mathbf{V}_k^*, u_k^*$ be an arbitrary optimal solution regarding (1). When running on the supporting hyperplane*

$$\mathbf{U}_k^{*\mathrm{T}}\mathbf{y} + u_k^* - \mathbf{V}_k^{*\mathrm{T}}\mathbf{x} = 0$$

via the trajectory

$$\mathbf{U}_k^{*\mathrm{T}}(1+\varepsilon_k)\mathbf{y}_k + u_k^* - \mathbf{V}_k^{*\mathrm{T}}(1+\delta)\mathbf{x}_k = 0,$$

the corresponding scale elasticity $\frac{\varepsilon_k}{\delta}$ depends on the actual output level.

Proof. The proof is an immediate consequence of Lemma 1 and Eq. (2), respectively. □

Theorem 1 postulates the reason for the misbehavior of the scale elasticity in the case of DRS; the following corollary summarizes this result, see Dellnitz and Rödder (2021):

Corollary 1. *Let $g_k^*, \mathbf{U}_k^*, \mathbf{V}_k^*, u_k^*$ be an arbitrary optimal solution regarding (1), we then have:*

- *for $u_k^* > 0$ and radially increasing outputs, $\frac{\varepsilon_k}{\delta}$ decreases.*
- *for $u_k^* < 0$ and radially increasing outputs, $\frac{\varepsilon_k}{\delta}$ increases.*

The issue of misleading values of $\frac{\varepsilon_k}{\delta}$ can thus be traced back to the fact that affine-linear supporting hyperplanes construct the facets of the technology or polyhedron. The following example illustrates the above statements.

Numerical Example

Consider the following BCC technology with one input and one output, as indicated in Fig. 5; the bold line indicates the efficient frontier, and the grey area is the feasible set. The DMUs or activities

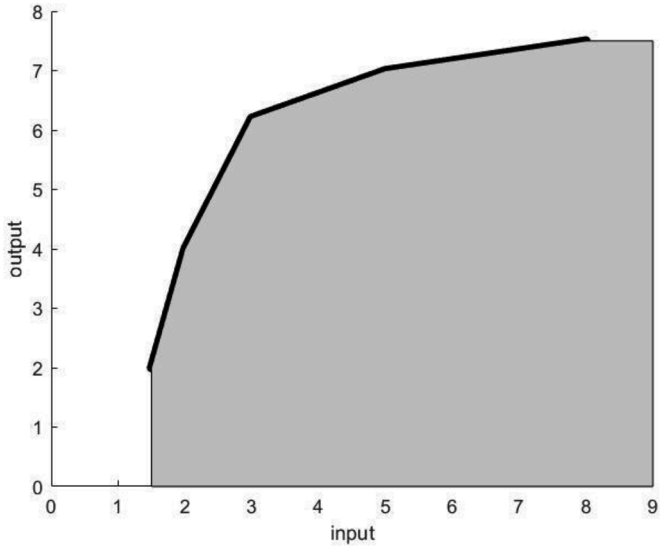

Fig. 5. BCC technology.

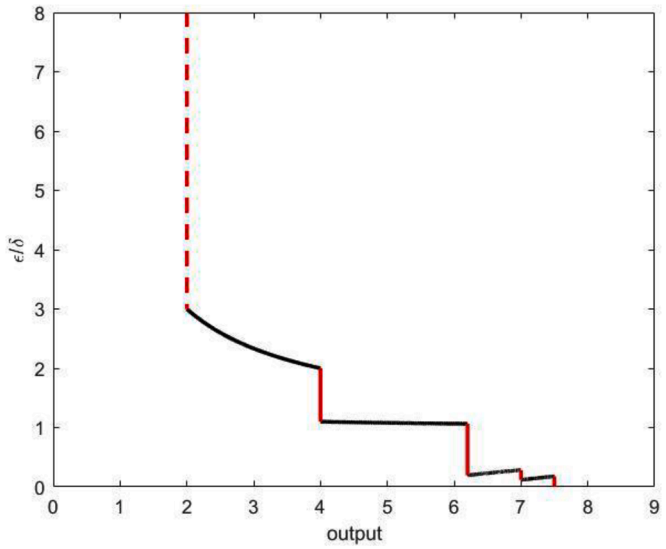

Fig. 6. Non-monotonicity of $\frac{\varepsilon}{\delta}$ in BCC.

determine the vertices of this technology: $(x_1,y_1) = (1.5,2)$; $(x_2,y_2) = (2,4)$; $(x_3,y_3) = (3,6.2)$; $(x_4,y_4) = (5,7)$; $(x_5,y_5) = (8,7.5)$. Fig. 6 shows the $\frac{\varepsilon}{\delta}$ values for the boundary of Fig. 5.

To approximate the graph in Fig. 6, we generate 6501 activities along the efficient frontier via incrementally increasing the input by 0.001 from 1.5 to 8 and calculating the corresponding efficient outputs. Next, we solve the optimization problem (3) for these efficient activities. The optimal solutions of the vertex activities $(x_1,y_1) = (1.5,2)$, $(x_2,y_2) = (2,4)$, $(x_3,y_3) = (3,6.2)$, $(x_4,y_4) = (5,7)$, $(x_5,y_5) = (8,7.5)$ are non-unique and thus $\frac{\varepsilon}{\delta}$ ditto. This non-uniqueness results in an interval with respect to $\frac{\varepsilon}{\delta}$ for each vertex activity leading to the red vertical lines in Fig. 6, the dashed line shall indicate that the $u_k^+$of (3) for the activity (x_1,y_1) is unbounded; hence, $\frac{\varepsilon}{\delta}$ goes to infinity. Table 1 presents the scale elasticity intervals.

The generated non-vertex activities on the efficient frontier have unique optimal solutions, and thus $\frac{\varepsilon}{\delta}$ is also unique, corresponding to the black parts of $\frac{\varepsilon}{\delta}$. In Fig. 6, we observe a monotonically decreasing behavior of $\frac{\varepsilon}{\delta}$ for the outputs between 2 and 6.2 (the IRS case), which is a consequence of the affine nature of the underlying hyperplanes and the fact that the intercept of these hyperplanes approaches the origin as one approaches MPSS located at $(x_3,y_3) = (3,6.2)$. This DMU is the only one that has an optimal solution with $\frac{\varepsilon}{\delta} = 1$, because $\frac{\varepsilon}{\delta} \in [0.194;\ 1.065]$, see Table 1. There is a nearly horizontal black line between outputs 4 and 6.2, but even here $\frac{\varepsilon}{\delta}$ still decreases monotonically, from 1.1 to 1.065; see the scale elasticity entries for DMU 2 and DMU 3 in Table 1.

The non-monotonicity is due to the increase in $\frac{\varepsilon}{\delta}$ between outputs 6.2 and 7, and between outputs 7 and 7.5; the DRS cases due to $\frac{\varepsilon}{\delta} < 1$. Worse still, Dellnitz and Rödder (2021) have recently shown that exploiting this non-monotonicity can generate very large activities with decreasing returns to scale, leading to a scale elasticity of close to one. These oversized activities thus lose their incentive to downsize.

These figures show that if we approach MPSS from the left, then $\frac{\varepsilon}{\delta}$

Table 1
Vertex activities and local scale elasticity intervals.

No	Input	Output	Scale elasticity intervals wrt (3)
1	1.5	2	3; ∞
2	2	4	1.1; 2
3	3	6.2	0.194; 1.065
4	5	7	0.119; 0.286
5	8	7.5	0; 0.178

decreases monotonically. However, when approaching this activity from the right, then $\frac{\varepsilon}{\delta}$ behaves non-monotonically. In other words, in the case of DRS, due to Eq. (2) composition, local scale elasticity increases rather than decreases as one moves from left to right along a supporting hyperplane (see Corollary 1). The next section studies this relationship in the context of multiplicative DEA.

4. Weak monotonicity of scale elasticities in multiplicative DEA

This section leads to a new result and shows that multiplicative DEA models do not suffer from the problem of non-monotonicity of local scale elasticities. For the sake of consistency, we immediately present the linearized dual formulation of the multiplicative DEA model:

$$\begin{aligned} \min\ & \dot{g}_k = \dot{\mathbf{U}}_k^{\mathrm{T}}\widetilde{\mathbf{y}}_k - \dot{\mathbf{V}}_k^{\mathrm{T}}\widetilde{\mathbf{x}}_k - \dot{u}_k \\ & \text{s.t.} \\ & \dot{\mathbf{U}}_k^{\mathrm{T}}\mathbf{1} = 1 \\ & \dot{\mathbf{U}}_k^{\mathrm{T}}\widetilde{\mathbf{y}}_j - \dot{\mathbf{V}}_k^{\mathrm{T}}\widetilde{\mathbf{x}}_j - \dot{u}_k \le 0 \qquad \forall j \in \mathscr{J} \\ & \dot{\mathbf{U}}_k, \dot{\mathbf{V}}_k \ge \mathbf{0} \text{ and } \dot{u}_k \text{ free} \end{aligned} \tag{4}$$

where the tilde sign denotes logarithms, and the dot indicates that the weights originate from a different technology. For detailed documentation on multiplicative models, their dual relationships, and respective properties, see Banker et al. (2004) and Zarepisheh et al. (2010).

To continue the reasoning consistently, we use the next lemma to develop the equations for determining scale elasticities in multiplicative models.

Lemma 2. *Let* $\dot{g}_k^*, \dot{\mathbf{U}}_k^* = \left(\dot{U}_{1k}^*, \cdots, \dot{U}_{sk}^*, \cdots \dot{U}_{Sk}^*\right)^{\mathrm{T}}, \dot{\mathbf{V}}_k^* = \left(\dot{V}_{1k}^*, \cdots, \dot{V}_{mk}^*, \cdots \dot{V}_{Mk}^*\right)^{\mathrm{T}}, \dot{u}_k^*$ *be an arbitrary optimal solution regarding* (4). *Then, radial output changes* $\mathbf{y}_k \to (1+\varepsilon_k)\mathbf{y}_k$ *under infinitesimal radial input changes* $\mathbf{x}_k \to (1+\delta)\mathbf{x}_k$ *must yield*

$$(1+\varepsilon_k) = (1+\delta)^{\sum_{m=1}^{M} \dot{V}_{mk}^*} \tag{5}$$

to maintain the (in)efficiency level of DMU k.

Proof To prove Lemma 2, first, we convert the log-linear efficiency equation by taking antilogarithms:

$$\prod_{s=1}^{S} y_{sk}^{\dot{U}_{sk}^*} = e^{\dot{u}_k^*} \prod_{m=1}^{M} x_{mk}^{\dot{V}_{mk}^*}$$

Embedding the scaling parameters $(1+\delta)$, $(1+\varepsilon_k)$, we obtain:

$$\prod_{s=1}^{S} \left[(1+\varepsilon_k)\cdot y_{sk}\right]^{\dot{U}_{sk}^*} = e^{\dot{u}_k^*} \prod_{m=1}^{M} \left[(1+\delta)\cdot x_{mk}\right]^{\dot{V}_{mk}^*}$$

$$\Rightarrow (1+\varepsilon_k)^{\sum_{s=1}^{S} \dot{U}_{sk}^*} \prod_{s=1}^{S} y_{sk}^{\dot{U}_{sk}^*} = (1+\delta)^{\sum_{m=1}^{M} \dot{V}_{mk}^*} e^{\dot{u}_k^*} \prod_{m=1}^{M} x_{mk}^{\dot{V}_{mk}^*}$$

$$\Rightarrow (1+\varepsilon_k)^{\underbrace{\sum_{s=1}^{S} \dot{U}_{sk}^*}_{=1}} = (1+\delta)^{\sum_{m=1}^{M} \dot{V}_{mk}^*} \underbrace{\frac{e^{\dot{u}_k^*} \prod_{m=1}^{M} x_{mk}^{\dot{V}_{mk}^*}}{\prod_{s=1}^{S} y_{sk}^{\dot{U}_{sk}^*}}}_{=1}$$

$$\Rightarrow (1+\varepsilon_k) = (1+\delta)^{\sum_{m=1}^{M} \dot{V}_{mk}^*} \qquad \square$$

For a different proof of Lemma 2, see Banker et al. (2004).

Now, logarithmization of $(1+\varepsilon_k) = (1+\delta)^{\sum_{m=1}^{M} \dot{V}_{mk}^*}$ leads to the scale elasticity estimate:

$$\frac{\ln(1+\varepsilon_k)}{\ln(1+\delta)} = \sum_{m=1}^{M} \dot{V}_{mk}^* \tag{6}$$

The following conclusions can now be drawn from Eqs. (5) and (6), see also Banker et al. (2004):

- Iff $\sum_{m=1}^{M} \dot{V}_{mk}^* > 1$ for all optimal solutions to (4) → increasing RTS (IRS)
- Iff $\sum_{m=1}^{M} \dot{V}_{mk}^* < 1$ for all optimal solutions to (4) → decreasing RTS (DRS)
- Iff $\sum_{m=1}^{M} \dot{V}_{mk}^* = 1$ for some optimal solutions to (4) → constant RTS (CRS)

To check for the above reasoning, one can solve (7):

$$\begin{aligned} \dot{v}_k^- = \inf\ & \dot{\mathbf{V}}_k^{\mathrm{T}}\mathbf{1} \text{ and } \dot{v}_k^+ = \sup\ \dot{\mathbf{V}}_k^{\mathrm{T}}\mathbf{1} \\ & \text{s.t.} \\ & \dot{\mathbf{U}}_k^{\mathrm{T}}\mathbf{1} = 1 \\ & \dot{\mathbf{U}}_k^{\mathrm{T}}\widetilde{\mathbf{y}}_j - \dot{\mathbf{V}}_k^{\mathrm{T}}\widetilde{\mathbf{x}}_j - \dot{u}_k \le 0 \qquad \forall j \in \mathscr{J}\backslash\{k\} \\ & \dot{\mathbf{U}}_k^{\mathrm{T}}\widetilde{\mathbf{y}}_k - \dot{\mathbf{V}}_k^{\mathrm{T}}\widetilde{\mathbf{x}}_k - \dot{u}_k = 0 \\ & \dot{\mathbf{U}}_k, \dot{\mathbf{V}}_k \ge \mathbf{0} \text{ and } \dot{u}_k \text{ free} \end{aligned} \tag{7}$$

Now, applying (7) with $\dot{v}_k^-$ and $\dot{v}_k^+$ being respective optimal objective function values, we then have:

- $\dot{v}_k^+ \ge \dot{v}_k^- > 1$ → increasing RTS (IRS)
- $\dot{v}_k^- \le \dot{v}_k^+ < 1$ → decreasing RTS (DRS)
- $\dot{v}_k^+ \ge 1 \ge \dot{v}_k^-$ → constant RTS (CRS)

With $\mathbf{y}_k \to (1+\varepsilon_k)\mathbf{y}_k$ and $\mathbf{x}_k \to (1+\delta)\mathbf{x}_k$, we form a trajectory on the corresponding supporting hyperplane $\prod_{s=1}^{S} y_s^{\dot{U}_{sk}^*} - e^{\dot{u}_k^*} \prod_{m=1}^{M} x_m^{\dot{V}_{mk}^*} = 0$, where the index k is omitted for the outputs and inputs to indicate their free variation. As a consequence, we can derive the following Theorem 2.

Theorem 2. *Let* $\dot{g}_k^*, \dot{\mathbf{U}}_k^*, \dot{\mathbf{V}}_k^*, \dot{u}_k^*$ *be an arbitrary optimal solution regarding* (4). *When running on the supporting hyperplane*

$$\prod_{s=1}^{S} y_s^{\dot{U}_{sk}^*} = e^{\dot{u}_k^*} \prod_{m=1}^{M} x_m^{\dot{V}_{mk}^*}$$

via the trajectory

$$\prod_{s=1}^{S} \left[(1+\varepsilon_k)\cdot y_{sk}\right]^{\dot{U}_{sk}^*} = e^{\dot{u}_k^*} \prod_{m=1}^{M} \left[(1+\delta)\cdot x_{mk}\right]^{\dot{V}_{mk}^*},$$

the corresponding scale elasticity is independent of the output level.

Proof. The proof follows from Lemma 2 and Eq. (5). $\square$

Theorem 2 illustrates that in the case of multiplicative models, the non-monotonic behavior of scale elasticities—as observed for BCC technologies—does not occur. Rather, it can be concluded that the output independence of Eq. (5) leads to constant scale elasticities on a trajectory, as given above. Accordingly, scale elasticities must behave weakly monotonic along the technology's surface. We continue the example already introduced to illustrate this fact.

Numerical Example (continued)

Again, we consider the efficient activities $(x_1, y_1) = (1.5, 2)$; $(x_2, y_2) = (2,4)$; $(x_3, y_3) = (3,6.2)$; $(x_4, y_4) = (5,7)$; $(x_5, y_5) = (8,7.5)$. Fig. 7 shows the graph of the scale elasticity values obtained by Eqs. (6) and (7). As in the classical DEA case, the optimal solutions of the vertex activities $(x_1, y_1) = (1.5,2)$, $(x_2, y_2) = (2,4)$, $(x_3, y_3) = (3,6.2)$, $(x_4, y_4) = (5, 7)$, $(x_5, y_5) = (8, 7.5)$ are non-unique in the multiplicative model, and thus, the same holds for the corresponding scale elasticities $\frac{\ln(1+\varepsilon_k)}{\ln(1+\delta)} = \sum_{m=1}^{M} \dot{V}_{mk}^*$ given by (6). Fathoming the intervals of $\frac{\ln(1+\varepsilon_k)}{\ln(1+\delta)}$ via (7) for the five vertex activities leads to the red vertical lines in Fig. 7, and the red dashed line shall indicate that the upper value of (7) for $(x_1, y_1) = (1.5,$

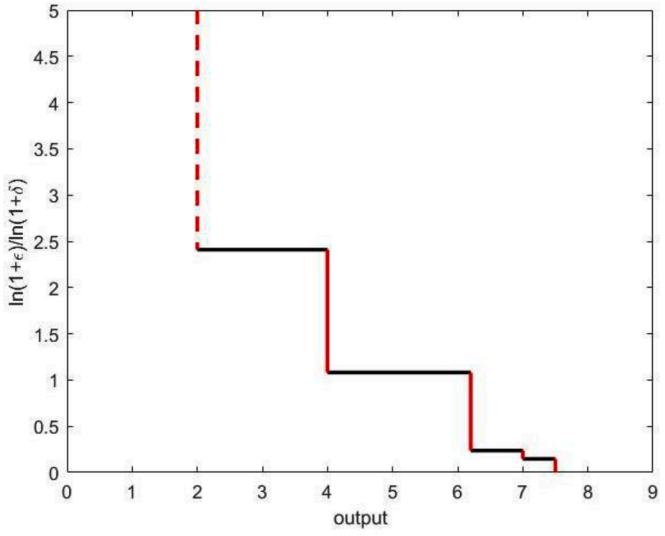

Fig. 7. Weak monotony of scale elasticity in multiplicative DEA.

Table 2
Vertex activities and scale elasticity intervals in multiplicative DEA.

No	Input	Output	Scale elasticity intervals wrt (7)
1	1.5	2	2.409; ∞
2	2	4	1.081; 2.409
3	3	6.2	0.238; 1.081
4	5	7	0.147; 0.238
5	8	7.5	0; 0.147

2) tends to infinity. Table 2 presents the scale elasticity intervals of the vertex activities, corresponding to the red lines in Fig. 7.

At first glance, the graph shown in Fig. 7 resembles the graph of scale elasticity presented in Fig. 6. The intervals of the vertex activities are very similar if one compares the scale elasticity intervals of the vertex activities in both models (see Tables 1 and 2). However, there is a significant difference: No increasing (black) trajectory exists. More precisely, the black parts between the vertical lines are flat. Fig. 7 demonstrates that scale elasticities satisfy the property of weak monotonicity in multiplicative DEA models.

An important economic consequence of the above observations is that we cannot—like in the classical BCC case, as demonstrated in Dellnitz and Rödder (2021)—generate a substantial activity beyond MPSS with decreasing returns to scale that pretends to have MPSS. So scale elasticities in multiplicative DEA models always point in the right direction to MPSS; i.e., the underlying technology and the local measure to substantiate scaling decisions are compatible. However, the decision-makers cannot gauge the distance to the MPSS due to the flat or horizontal lines.

Monotonicity of scale elasticities is a compelling property that must be satisfied. Otherwise, the measure does not accurately reflect scaling potentials, and this failure can imply disincentives. From this point of view, approximating a real technology via multiplicative models is preferable to classical BCC technologies. However, it would be desirable to have an adequate measure in other technologies, as discussed in the next section, that both fulfill the monotonicity condition and reveal the remaining distance to the MPSS.

5. Monotonicity via a global scaling index

In the first step, we have pointed out some irregularities, i.e., a non-monotonic behavior, of local scale elasticities caused by the affine-linear structure of the supporting hyperplanes (see Eq. (2) and Theorem 1). As shown in the previous section, this problem can be circumvented by changing the technology. From a scientific point of view, finding a measure that works correctly for most or even all technologies is desirable.

To fill this gap, we propose a new index, measuring the distance to a DMU exhibiting the most productive scale size (MPSS)—a DMU that is efficient in both the CCR and BCC worlds (e.g., Banker, 1984; Zhu & Shen, 1995; Fukuyama, 2003; Dellnitz et al., 2018). Consequently, our approach is conceptually related to global RTS (see Podinovski, 2004a, 2004b). For the first time, we prove that the approach implies monotonicity. The starting point is the following envelopment form of the classical CCR problem:

$$\begin{aligned} &\min\ h_k \\ &\text{s.t.} \\ &h_k\mathbf{x}_k - \alpha_k\sum_j \lambda_{kj}\mathbf{x}_j \geq 0 \\ &\qquad \alpha_k\sum_j \lambda_{kj}\mathbf{y}_j \geq \mathbf{y}_k \\ &\qquad \sum_j \lambda_{kj} = 1 \\ &\lambda_{kj} \geq 0\ \forall j \in \mathscr{J} \text{ and } \alpha_k > 0 \end{aligned} \tag{8}$$

In Eq. (8), for the sake of transparency, we omit the case of partial inefficiencies, but it can be easily inserted if needed. Eq. (8) is the nonlinear version of the CCR problem due to the scaling parameter α. However, this CCR problem shows the convexity constraint, which is likewise present in the CCR world and is important for our development. When substituting $\lambda'_{kj} = \alpha_k\lambda_{kj}$ in Eq (8), one obtains the linear problem (9).

$$\begin{aligned} &\min\ h_k \\ &\text{s.t.} \\ &h_k\mathbf{x}_k - \sum_j \lambda'_{kj}\mathbf{x}_j \geq 0 \\ &\qquad \sum_j \lambda'_{kj}\mathbf{y}_j \geq \mathbf{y}_k \\ &\qquad \sum_j \lambda'_{kj} = \alpha_k \\ &\lambda'_{kj} \geq 0\ \forall j \in \mathscr{J} \text{ and } \alpha_k > 0 \end{aligned} \tag{9}$$

With this transformation, the convexity constraint is softened by α_k, and thus it does not have a restrictive effect when minimizing the efficiency factor h_k. Eq (9) allows determining a path to MPSS, as demonstrated in Banker (1984), Cooper et al. (1996), and Esfandiar et al. (2022):

Corollary 2. *Let h_k^*, $\lambda_{kj}^{'*}$ and $\sum_j \lambda_{kj}^{'*} = \alpha_k^*$ be an arbitrary optimal solution regarding (9). Then, we have the following statements:*

1. *Iff $h_k^* = 1$ holds and all slacks are zero, then DMU k has MPSS; see Banker (1984) for proof of this.*
2. *If $h_k^* < 1$ and all slacks are zero, then DMU k can be transformed into an MPSS by $\left(\frac{h_k^*}{\alpha_k^*}\mathbf{x}_k, \frac{1}{\alpha_k^*}\mathbf{y}_k\right)$.* Here, the scaling direction depends on α_k^*; i.e., for $\alpha_k^* < 1$, it is an upscaling (increasing RTS), and for $\alpha_k^* > 1$, it is a downscaling (decreasing RTS). This preliminary work now allows for the definition of the global scaling index.

Definition 1. *Let h_k^* and $\sum_j \lambda_{kj}^{'*} = \alpha_k^*$ be an arbitrary optimal solution regarding* (9). *Then, we call*

$$\varphi_k^* = \frac{1}{\alpha_k^*} \tag{10}$$

the global scaling index of DMU k.

It is important to mention that if the slacks are nonzero after scaling

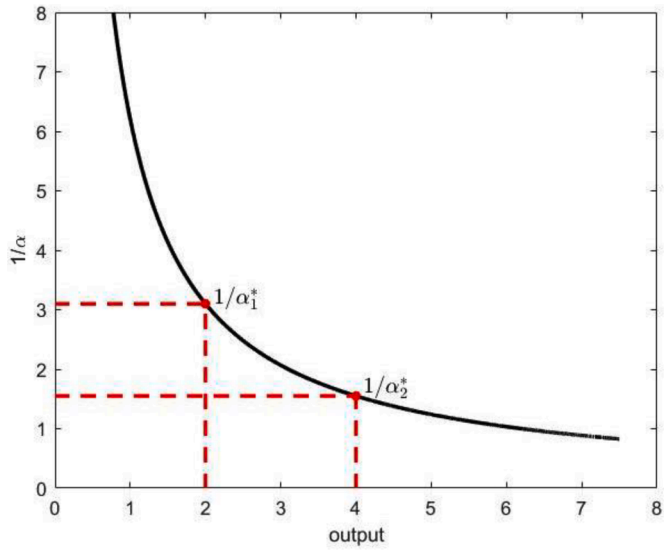

Fig. 8. Monotonicity of $\frac{1}{\alpha}$ in BCC.

the activity, contrary to the condition in Corollary 2, we need to adjust the input and output vectors according to these slacks to make the activity an MPSS; for more details on correcting for partial inefficiencies or slacks, see Cooper et al. (1996) and Esfandiar et al. (2022). The latter point, however, does not diminish the managerial implications and (weak) monotonicity property of the global scaling index, which has yet to be proven since the principle of proportional scaling as defined by the second statement of Corollary 2—which is also the philosophy of the flawed concept of local scale elasticity—remains valid.

From a geometric point of view, we illustrate the quality of the global scaling index and verify that for a BCC technology, it satisfies the property of (weak) monotonicity—sometimes even in a strict form.

Numerical Example (continued)

Again, we consider the activities $(x_1,y_1) = (1.5,2)$; $(x_2,y_2) = (2,4)$; $(x_3,y_3) = (3,6.2)$; $(x_4,y_4) = (5,7)$; $(x_5,y_5) = (8,7.5)$.

Now, we solve Eqs. (9) and (10) for the (weakly efficient) activities of the boundary of BCC technology, leading to the graph depicted in Fig. 8. Fig. 8 shows the global scaling index—i.e., the values for $\frac{1}{\alpha}$ on the full BCC technology. Interestingly, the graph is strictly monotonically decreasing and thus exhibits the desired property. We have also marked

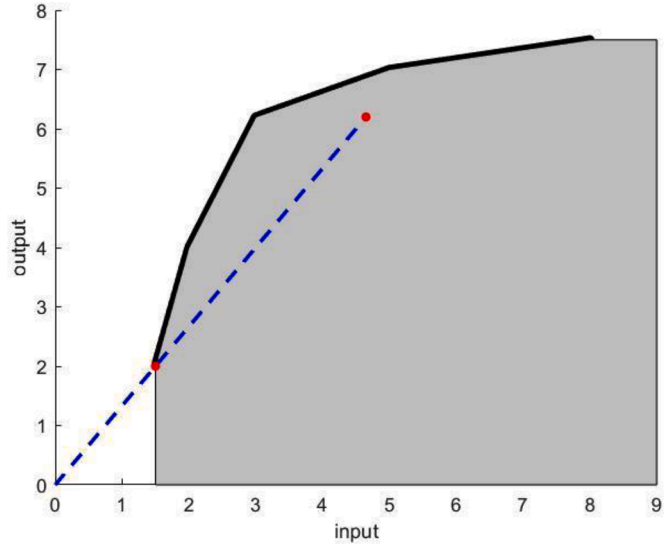

Fig. 9. Illustration of $\frac{1}{\alpha}$ for activity 1.

the two scaling indices $\frac{1}{\alpha}$ for activity 1 and activity 2 in red. Here, $\frac{1}{\alpha_1^*} = 3.1$ means that activity $(x_1,y_1) = (1.5,2)$ must increase its activity from $(1.5,2)$ to $(4.65,6.2) = (1.5,2)\cdot 3.1$ to reach MPSS level; the activity $(x_2,y_2) = (2,4)$ is closer to the MPSS unit and consequently $\frac{1}{\alpha_2^*} = 1.55$ is also smaller.

Fig. 9 illustrates the scaling path for the activity $(x_1,y_1) = (1.5,2)$ with $\frac{1}{\alpha_1^*} = 3.1$; that is, the scaling from the first red point $(1.5,2)$ to the second $(4.65,6.2) = (1.5,2)\cdot 3.1$.

The theoretical explanations and the last example have shown that the new measure fulfills the desired properties; more surprisingly, the monotonicity property in the example was even strictly fulfilled. However, this is not always true: When multiple MPSS are present in a BCC technology—even with a single input and single output—the global scaling index decreases monotonically in a weak sense.

The MPSS scaling path, as given in the second statement of Corollary 2, is not necessarily unique, nor is the global scaling index. Therefore, Banker et al. (1996) propose the following optimization problem to fathom the interval in which $\sum_j \lambda_{kj}^{'}$ can vary:

$$
\begin{aligned}
&\alpha_k^- = \min \alpha_k \text{ and } \alpha_k^+ = \max \alpha_k \\
&\text{s.t.} \\
&h_k^* \mathbf{x}_k - \sum_j \lambda_{kj}^{'} \mathbf{x}_j \geq 0 \\
&\sum_j \lambda_{kj}^{'} \mathbf{y}_j \geq \mathbf{y}_k \\
&\sum_j \lambda_{kj}^{'} = \alpha_k \\
&\lambda_{kj}^{'} \geq 0 \ \forall j \in \mathcal{J} \text{ and } \alpha_k > 0
\end{aligned}
\tag{11}
$$

Corollary 3. *Let $h_k^* \leq 1$ and α_k^-, α_k^+ be the optimal solutions to Eq. (10). To determine the minimal scaling path towards MPSS, the following rationale holds:*

- $1 < \frac{1}{\alpha_k^-} \leq \frac{1}{\alpha_k} \rightarrow$ *increasing RTS (IRS); where* $\frac{1}{\alpha_k^\star} := \frac{1}{\alpha_k^-}$ *determines the minimal scaling path.*
- $1 > \frac{1}{\alpha_k^-} \geq \frac{1}{\alpha_k^+} \rightarrow$ *decreasing RTS (DRS); where* $\frac{1}{\alpha_k^\star} := \frac{1}{\alpha_k^-}$ *determines the minimal scaling path.*
- $\frac{1}{\alpha_k^-} \geq 1 \geq \frac{1}{\alpha_k^+} \rightarrow$ *constant RTS (CRS); where* $\frac{1}{\alpha_k^\star} := 1$ *because DMU k already has the right scale size but only has to improve efficiency or eliminate partial inefficiencies, if necessary.*

We use the rationale of Corollary 3 to assign a unique scaling index to the activities under consideration; this reasoning is compatible with the minimum principle often propagated in economics.

To demonstrate the effect or impact of multiple MPSS activities on the global scaling index $\frac{1}{\alpha}$, we study a modified numerical example.

Numerical Example (continued)

We consider the BCC efficient activities $(x_1,y_1) = (1.1,1.1)$; $(x_2,y_2) = (2,3)$; $(x_3,y_3) = (3,4.5)$; $(x_4,y_4) = (5,5)$. These activities lead to a modified BCC technology, as shown in Fig. 10. The activities $(x_2,y_2) = (2,3)$ and $(x_3,y_3) = (3,4.5)$ are both MPSS, but so is any activity that can be generated by convex combinations $\lambda(x_2,y_2) + (1-\lambda)(x_3,y_3)$. As a consequence, every activity which is located on the red line in Fig. 10 is BCC and CCR efficient, and hence, we can find an optimal solution to Eq. (10) for it with $h_k^* = 1$ and $\frac{1}{\alpha_k^\star} = 1$.

The latter statement and the rationale of Corollary 3 make $\frac{1}{\alpha}$ a weakly monotonically decreasing index, which is illustrated in Fig. 11. Here, the range for optimal solutions of Eq. (10), i.e., $\frac{1}{\alpha_k^-}$, $\frac{1}{\alpha_k^+}$ is depicted by the shaded area. The blue and the red line, which is connected, show the

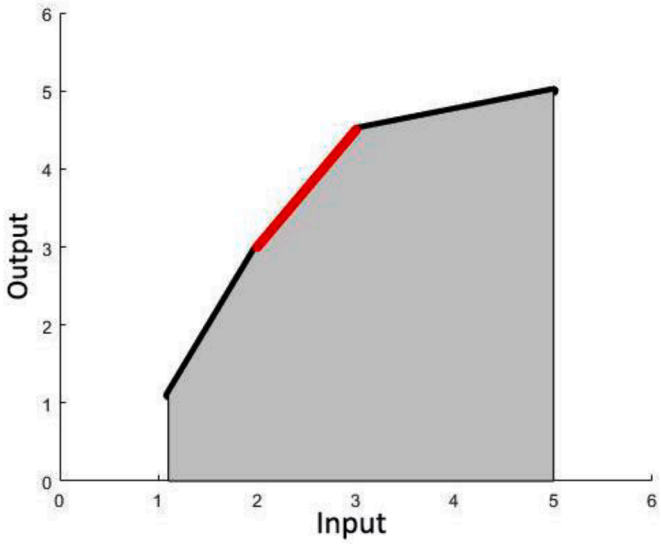

Fig. 10. Modified BCC technology.

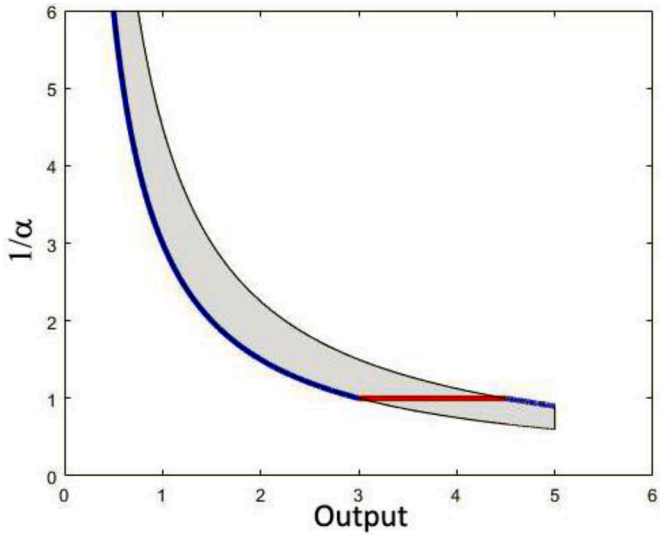

Fig. 11. Weak monotonicity of $\frac{1}{\alpha}$ in BCC.

curve we obtain when using the rationale of Corollary 3. This is the weakly monotonic function we are looking for.

The—more intuitively gained—insights are now to be formally substantiated. First, we give:

Lemma 3. *Let* $h_k^* < 1$ *and* $\sum\lambda'^*_{kj} \in [\alpha_k^-, \alpha_k^+]$ *be an optimal solution regarding* (9). *Then, we get that* $\sum_j \lambda'^*_{kj} \in [\alpha_k^-, \alpha_k^+]$ *is independent with respect to further proportional disposal of the inputs of* $(\mathbf{x}_k,\ \mathbf{y}_k)$.

Proof. According to the requirements, we have that $h_k^* < 1$ and $\sum\lambda'^*_{kj} = \alpha_k^* \in [\alpha_k^-, \alpha_k^+]$ is an optimal solution to Eq. (9), satisfying the constraints

$$h_k^*\mathbf{x}_k - \sum_j \lambda'^*_{kj}\,\mathbf{x}_j \geq 0$$
$$\sum_j \lambda'^*_{kj}\,\mathbf{y}_j \geq \mathbf{y}_k$$
$$\sum_j \lambda'^*_{kj} = \alpha_k^*$$

with some restrictions being active, of course.

Now, we construct an activity $\mathbf{x}_l$ meeting the equation $\mathbf{x}_l = t{\cdot}\mathbf{x}_k$, where $t > 1$ is a scalar that can arbitrarily be chosen. Next, one has to replace $\mathbf{x}_k$ in the above restrictions, obtaining:

$$h_k^*\cdot\frac{1}{t}\cdot\mathbf{x}_l - \sum_j \lambda'^*_{kj}\,\mathbf{x}_j \geq 0$$
$$\sum_j \lambda'^*_{kj}\,\mathbf{y}_j \geq \mathbf{y}_k$$
$$\sum_j \lambda'^*_{kj} = \alpha_k^*$$

Compressing $h_k^*{\cdot}1/t$ to h_l^*, we have an optimal solution to Eq. (9) for $(\mathbf{x}_l, \mathbf{y}_k)$. Supposing now that h_l^* and $\sum\lambda'^*_{kj} = \alpha_k^*$ are suboptimal to Eq. (9) leads to a contradiction because we can simply reverse this transformation. That is, the suboptimality of h_l^* and $\sum\lambda'^*_{kj} = \alpha_k^*$ regarding Eq. (9) contradicts the optimality of h_k^* and $\sum\lambda'^*_{kj} = \alpha_k^*$ w.r.t. Eq. (9). □

Lemma 3 makes the proof of the next theorem easier accessible because radial disposals of inputs do not additionally dilute the properties of the global scaling index.

Theorem 3. *Applying the rationale of Corollary 3,* $\frac{1}{\alpha}$ *behaves weakly monotonically on the entire BCC technology.*

Proof. To prove this theorem, w.l.o.g. we can pick any feasible BCC efficient but CCR inefficient activity $(\mathbf{x}, \mathbf{y})$; i.e., for $(\mathbf{x}, \mathbf{y})$, we have $g^* = 1$ and $h^* < 1$. Let $\frac{1}{\alpha^\star}$ be the corresponding global scaling index according to Corollary 3. We have to prove the theorem by two cases, namely for $\frac{1}{\alpha^\star} > 1$ (situation of increasing RTS) and $\frac{1}{\alpha^\star} < 1$ (situation of decreasing RTS); we show only the reasoning for $\frac{1}{\alpha^\star} > 1$, the other case can be derived in the same way.

First, suppose we have $\frac{1}{\alpha^\star} > 1$ (situation of increasing RTS). Now, when choosing $\frac{1}{\alpha}$ such that $\frac{1}{\alpha^\star} > \frac{1}{\alpha} > 1$ and scaling $(\mathbf{x}, \mathbf{y})$ up via $(\mathbf{x}', \mathbf{y}') = \left(\frac{1}{\alpha}\mathbf{x}, \frac{1}{\alpha}\mathbf{y}\right)$, then $(\mathbf{x}', \mathbf{y}')$ has the same CCR efficiency h^* as $(\mathbf{x}, \mathbf{y})$ does. This transformation must lead to a feasible activity due to $\frac{1}{\alpha^\star} > \frac{1}{\alpha} > 1$. Next, it follows that the new global scaling index $\frac{1}{\alpha'^\star}$ of $(\mathbf{x}', \mathbf{y}')$ to reach MPSS must be smaller than $\frac{1}{\alpha^\star}$ of $(\mathbf{x}, \mathbf{y})$ because of $(\mathbf{x}', \mathbf{y}') > (\mathbf{x}, \mathbf{y})$, if all inputs/outputs are nonzero, and the new global scaling index can directly be calculated via $\frac{1}{\alpha'^\star} = \frac{\frac{1}{\alpha^\star}}{\frac{1}{\alpha}}$. Due to the proportional scaling of inputs and outputs, which is only a CCR invariant transformation, we have $g'^* < g^* = 1$ for the BCC efficiencies, which can be checked by applying Eq. (1). Projecting $(\mathbf{x}', \mathbf{y}')$ onto the boundary via $(\hat{\mathbf{x}}' := g'^*\mathbf{x}', \mathbf{y}')$, ultimately, one can make use of Lemma 3 and realizes that $\frac{1}{\alpha'^\star}$ must also be optimal for $(\hat{\mathbf{x}}', \mathbf{y}')$.

For the case $\frac{1}{\alpha^\star} < 1$ (situation of decreasing RTS), one can make use of nearly the same lines of argumentation. The only difference is that we need to reverse some relations as we are now dealing with downsizing processes.

The global scaling index proposed in Corollary 3 satisfies the property of weak monotonicity since we can apply this inference process to any arbitrary (feasible) activity of the BCC technology. □

The latter theorem shows that the global scaling index always works weakly monotonic because it measures the distance to an MPSS activity—a global reference point from a technological perspective. Due to this feature, we can apply this concept to several, if not all, technologies, e.g., free disposal hull, multiplicative, etc. Hence, the above general property makes the global scaling index more powerful than local scale elasticities. To illustrate that the global scaling index addresses the weaknesses of the concept of local scale elasticities, we

consider an application from the banking sector in the next section.

6. The global scaling index: a better tool to drive banks' scale decisions

Dellnitz and Rödder (2021) demonstrate the misleading nature of local scale elasticities via a dataset of 37 Brazilian banks studied first by Henriques et al. (2018). In this context, the authors consider three inputs (fixed assets, total deposits, and personnel expenses) and one output (total loans), making use of the so-called intermediation approach (Henriques et al., 2018). Accordingly, Table 3 presents the data for the business year 2016 in thousands of Brazilian reais.

We obtain the solutions in Table 4 when solving the optimization problems presented for the 37 Brazilian banks. Four banks are of MPSS: Alfa, Cooperativo Sicredi, Ribeirão Preto, and Intermedium—these banks are efficient in all three models, and the corresponding CCR and BCC multipliers are nonzero. Different from Esfandiar et al. (2022), we do not need to normalize our data because we are only interested in determining the (minimal) scaling path to an MPSS, not in the partial inefficiencies—which is the next step after scaling a DMU to the MPSS level. Consequently, the optimization problems remain units invariant (Lovell & Pastor, 1995), so our considerations are not affected by the magnitude of the data. The fifth column of Table 4 shows the local scale elasticities determined by Eqs. (2) and (3). The most interesting observation is that the DMUs 23 and 32 operate under DRS, and their local scale elasticities draw near one—usually indicating that they are close to an MPSS unit. The latter observation is a consequence of non-monotonicity. When checking the local scale elasticities of both DMUs in the sixth column, obtained by the multiplicative model, the incentive effect of reducing activity is noticeably greater due to the more significant productivity gain. However, DMUs 23 and 32 display the same productivity gains due to the measure's weak monotonicity property. The last column shows the enormous distance from MPSS activity for both banks, but Original (DMU 32) is nevertheless significantly closer to MPSS than BMG (DMU 23). DMU 9 (Arbi), for example, has a low level of loans but a high level of deposits, fixed assets, and personnel expenses compared to DMU 22 (Ribeirão Preto). Here, the unique $\frac{1}{\alpha}$ indicates that Arbi should increase its activity by approximately 8.41 to reach an MPSS level; additionally, it should significantly improve its efficiency level (CCR and BCC). Such an activity redesign involves a huge effort and, thus, the recommendation according to $\frac{1}{\alpha}$ is strategic.

The information on the scale elasticities is only local and does not show the long path Arbi, for example, has ahead of it in realizing this growth. There are various possibilities here: internal or external

Table 3
Activities of 37 Brazilian banks.

Bank	No	Fixed assets	Total deposits	Personnel expenses	Total loans
Alfa	1	323,417	40,626	309,151	6748,462
Bonsucesso	2	301,688	8779	939,761	308,364
Semear	3	1689	3023	567,958	400,229
Topázio	4	4005	3197	266,165	147,256
Banestes	5	276,560	79,967	9310,156	3473,396
Banif	6	6696	6029	490,918	73,687
Banrisul	7	956,272	401,681	37,793,700	29,808,188
BB	8	31,221,063	5246,319	455,560,520	667,786,191
Arbi	9	8558	1544	70,717	46,319
Capital	10	354	657	5478	3079
Cooperativo Sicredi	11	151,596	26,463	10,362,623	14,442,009
Banco da Amazônia	12	278,514	130,794	2909,788	3873,265
Banco da China Brasil	13	6451	4777	294,503	484,293
Banese	14	82,376	39,238	2895,553	2050,738
Banpará	15	114,978	67,197	3884,973	3431,025
BNB	16	236,206	426,027	10,352,508	12,678,428
Fibra	17	78,659	23,233	2173,689	2479,147
Ficsa	18	1074	972	79,236	6116
La Nacion Argentina	19	16,351	1251	4433	29,052
Luso Brasileiro	20	12,463	5876	639,616	697,948
Rep Oriental Uruguay	21	2294	553	1272	14,248
Ribeirão Preto	22	1575	1698	67,483	373,867
BMG	23	1873,997	46,798	5200,705	8087,786
Bradesco	24	51,076,723	3209,178	189,864,277	317,809,283
BRB	25	418,334	214,699	9157,803	9522,840
CEF	26	13,153,796	5018,876	451,018,737	672,513,474
Citibank	27	619,525	296,551	14,677,936	16,009,264
HSBC	28	3099,668	894,990	55,709,668	55,630,103
Intermedium	29	6627	14,391	1220,503	2187,713
Itaú	30	84,219,449	3641,920	297,347,284	396,500,032
Mercantil do Brasil	31	235,083	87,432	7825,089	7646,678
Original	32	728,170	35,671	1466,660	2587,370
Panamericano	33	840,450	87,330	12,960,426	16,230,243
Rendimento	34	38,449	29,799	583,234	318,071
Safra	35	3099,710	440,788	9228,824	38,610,052
Santander	36	16,448,887	1736,403	137,822,766	212,243,750
Sofisa	37	83,495	16,278	2885,708	1738,000

Table 4
Efficiencies, local scale elasticities, and global scaling indexes.

No	CCR eff.	BCC eff.	Mult. eff	$\frac{\varepsilon_k^-}{\delta}; \frac{\varepsilon_k^+}{\delta}$	$\dot{v}_k^-; \dot{v}_k^+$	$\frac{1}{\alpha_k^+}; \frac{1}{\alpha_k}$
1	1.00	1.00	1.00	0.1636; 1.0104	0.6316; 1.2517	1.0000
2	0.12	0.15	0.10	1.0470	1.0725	1.8945
3	0.80	0.91	0.98	1.2587	1.2295	2.3257
4	0.19	0.31	0.18	1.2272	0.9933	2.9822
5	0.15	0.19	0.16	0.9878	0.8659	0.1788
6	0.05	0.13	0.05	1.1110	0.8616	5.5581
7	0.27	0.53	0.40	0.9751	0.8659	0.0180
8	0.51	0.98	0.99	0.9193	0.0000	0.0024
9	0.13	0.40	0.16	1.5188	2.9124	8.4126
10	0.09	1.00	1.00	12.0746; ∞	2.5518; 6.0000	140.2599
11	1.00	1.00	1.00	0.2417; 1.0203	0.7323; 1.3305	1.0000
12	0.20	0.39	0.24	0.9114	0.8402	0.1234
13	0.42	0.46	0.41	0.7395	0.8616	0.8952
14	0.21	0.32	0.23	0.9651	0.8659	0.2292
15	0.22	0.43	0.28	0.8881	0.8659	0.1232
16	0.23	0.79	0.50	0.9514	0.8395	0.0297
17	0.40	0.52	0.35	0.9539	0.8659	0.2158
18	0.03	0.66	0.16	2.8118	4.4120	66.7775
19	0.30	0.51	0.55	1.5595	1.2517	232.2891
20	0.42	0.44	0.36	0.8385	1.0725	0.8469
21	0.51	1.00	1.00	1.9527; ∞	1.1215; 6.0000	473.6428
22	1.00	1.00	1.00	0.3479; 1.5694	0.8395; 4.4120	1.0000
23	0.60	0.64	0.50	0.9968	0.8112	0.0742
24	0.42	1.00	0.82	0.8180; 0.9431	0.6316	0.0013
25	0.20	0.51	0.31	0.9566	0.8402	0.0401
26	0.51	1.00	1.00	0.0000; 0.9979	0.0000; 0.8659	0.0009; 0.0007
27	0.24	0.59	0.36	0.9475	0.8402	0.0250
28	0.26	0.55	0.38	0.9594	0.8659	0.0078
29	1.00	1.00	1.00	0.2561; 1.0549	0.7544; 1.2295	1.0000
30	0.42	0.86	0.79	0.9602	0.8112	0.0012
31	0.33	0.55	0.38	0.9414	0.8659	0.0681
32	0.33	0.36	0.27	0.9944	0.8112	0.1459
33	0.58	0.75	0.56	0.8024	0.8112	0.0472
34	0.09	0.09	0.09	1.0075	0.8402	1.3637
35	0.46	1.00	0.67	0.3695; 0.8974	0.6316	0.0189
36	0.48	0.96	0.77	0.9798	0.8112	0.0023
37	0.31	0.31	0.26	0.9927	1.0725	0.5352

growth—e.g., Arbi could consider entering into a strategic alliance with the smaller DMUs (e.g., 10, 18, 19, etc.) to realize some of the growth. This also shows one of the weaknesses of the new index, namely that it only determines the scaling path to maximum productivity but not the optimal sub-decisions to realize it; future work could remedy this shortcoming by developing an interactive tool.

A similar reasoning applies to the DMUs 10, 18, 19, and 21, but the required level of activity scaling is even higher. Likewise, the large numbers could hint that Henriques et al. (2018) compared inhomogenous DMUs—banks with different business models or asset foci; local scale elasticities do not encourage such thoughts.

The shortcomings discussed in this manuscript are particularly evident in DMUs 13 and 23. For example, DMU 13 has a scale efficiency (CCR divided by BCC efficiency) of 0.91 and a scale elasticity of 0.7395 (DRS); DMU 23, on the other hand, has a scale efficiency of around 0.94 and a scale elasticity of almost 1 (0.9968)—the conclusion that 23 is closer to the maximum productivity is nevertheless inadmissible: Downscaling from DMU 13 can be realized with 0.8952; downsizing from DMU 23, on the other hand, requires a factor of 0.0742! This means that the latter global index implies that DMU 23 must make much more effort to reach the productivity maximum if it intends to achieve this.

The above observations demonstrate the global scaling index is more suitable than the local scale elasticities in BCC technologies because of its monotonicity property. This measure reveals—even in piecewise linear technologies—the direction and distance to MPSS activities and thus more reliably assists management in improving productivity.

7. Economic implications

From a conceptual and economic perspective, elasticities have been a standard tool of micro- and macroeconomic theory since Alfred Marshall (1885)—i.e., for more than one hundred years. Here, they are used to motivate decisions on substitutions, expansions, or contractions etc. by organizations or institutions; in other words, they are, among other things, a standard tool for classifying the size of companies, regions, or countries as evidenced by myriads of publications; a quick search on the Web of Science returns over 17,000 publications on elasticities.

About 40 years ago, Banker et al. (1984) transferred this idea to activities comprising multiple inputs and multiple outputs—here, too, with a rigorous conceptual approach. However, in their paper and the subsequent analyses, it was not noticed that the approximation of production functions via DEA using piecewise linear functions undermines the well-grounded concept of scale elasticities. It might not have been noticed over the years because the issue resolved in this paper cannot happen with smooth and differentiable functions—as they were always assumed in earlier micro- and macroeconomic considerations (e.g., Cobb-Douglas production functions).

This means that analyses via DEA can be noisy when scale elasticities of organizations or institutions are applied to determine increasing, constant, or decreasing returns to scale. This observation suggests two basic attitudes:

First, we may reject the BCC instrument outright, as the method may imply misclassifications. So we could, for example, prefer piecewise Cobb-Douglas production functions as the technology; see Section 4 again. However, this is not very desirable from a practical point of view due to the success of the classical BCC technology in DEA-based studies.

Second, we can try to find a way to avoid these misclassifications in the BCC model. Our global index provides such a corrective; cf. Section 5.

Economic theories have always been characterized by diversity in their conceptual approaches and the associated problem solutions; consequently, our idea may only be the beginning of developing new model adaptations or indices that resolve the shortcomings discussed in this manuscript.

8. Conclusions and future research directions

Improving productivity is increasingly important in the DEA literature. In this context, the concept of local scale elasticities arising from technologies with VRS should assist an analyst in inferring respective potentials. The concept of local scale elasticities resulting, however, has been recently questioned in the literature. This study shows that in classical VRS technologies—enveloped by affine-linear supporting hyperplanes—local scale elasticities can blur sizing potentials: the more, the larger a company is. This is due to the non-monotonicity of local scale elasticities. We also show this issue does not apply to multiplicative DEA models. Here, local scale elasticities demonstrate a weakly monotonic behavior. We propose a global scaling index with the desired property in most if not all, technologies. The new index is shown to be (weakly) monotonic and, thus, a more reliable candidate for incentivizing scaling decisions than local scale elasticities. Ultimately, we exhibit its practical usefulness via an application stemming from the Brazilian banking system. In particular, strategic decisions about the scale of operations should be made based on the new global scaling index rather than on local scale elasticities because of the shortcomings of local scale elasticities. Large or high-volume companies are common in our modern business world, especially in banking. When DEA analysts here rely on traditional indices, scaling potential may go unrecognized, and banks may be reported as highly productive even though the scale of operations is not optimally designed. The new index reveals a company's weakness is due to its properties.

Further studies on adequately embedding local and global scaling indicators in incentive schemes—e.g., in the context of centralized management—might be a worthwhile focus for future research. Due to its strategic nature, it would be fruitful to advance the presented global scaling index to a (time) robust indicator that prepares a DMU for future periods. Another exciting avenue might be to extend the idea of the new index to network models or pollution-generating technologies; especially the latter context could be of public interest due to the green movements in almost all developed countries, where scaling issues are currently hotly debated.

CRediT authorship contribution statement

Andreas Dellnitz: Conceptualization, Formal analysis, Methodology, Validation, Writing – original draft, Writing – review & editing. **Madjid Tavana:** Formal analysis, Methodology, Visualization, Writing – original draft, Writing – review & editing.

References

Asmild, M., Hollingsworth, B., & Birch, S. (2013). The scale of hospital production in different settings: One size does not fit all. *Journal of Productivity Analysis, 40*, 197–206.

Assani, S., Jiang, J. L., Cao, R. M., & Yang, F. (2018). Most productive scale size decomposition for multi-stage systems in data envelopment analysis. *Computers & Industrial Engineering, 120*, 279–287.

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). Some models for estimating technical and scale inefficiences in data envelopment analysis. *Management Science, 30*, 1078–1091.

Banker, R. D., Cooper, W. W., Seiford, L. M., Thrall, R. M., & Zhu, J. (2004). Returns to scale in different DEA models. *European Journal of Operational Research, 154*, 345–362.

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

Butler, T. W., & Li, L. (2005). The utility of returns to scale in DEA programming: An analysis of michigan rural hospitals. *European Journal of Operational Research, 161*, 469–477.

Charnes, A., Cooper, W. W., & Rhodes, E. (1978). Measuring the efficiency of decision making units. *European Journal of Operational Research, 2*, 429–444.

Cook, W. D., & Zhu, J. (2011). Multiple variable proportionality in data envelopment analysis. *Operations Research, 59*, 1024–1032.

Cooper, W. W., Thompson, R. G., & Thrall, R. M. (1996). Chapter 1 Introduction: Extensions and new developments in DEA. *Annals of Operations Research, 66*, 1–45.
Dellnitz, A., Kleine, A., & Rödder, W. (2018). CCR or BCC: What if we are in the wrong model? *Journal of Business Economics, 88*, 831–850.
Dellnitz, A., & Rödder, W. (2021). Returns to scale as an established scaling indicator: Always a good advisor? *Jahrbücher für Nationalökonomie und Statistik, 241*(2), 173–186.
Dellnitz, A., Tavana, M., & Banker, R. (2022). A novel median-based optimization model for eco-efficiency assessment in data envelopment analysis. *Annals of Operations Research*. https://doi.org/10.1007/s10479-022-04937-4
Esfandiar, E., Eslami, R., Khoveyni, M., & Gilani, A. (2022). Identifying the closest most productive scale size unit in data envelopment analysis. *OR Spectrum*. https://doi.org/10.1007/s00291-022-00692-x
Førsund, F. R. (1996). On the calculation of the scale elasticity in DEA models. *The Journal of Productivity Analysis, 7*, 283–302.
Fukuyama, H. (2000). Returns to scale and scale elasticity in data envelopment analysis. *European Journal of Operational Research, 125*, 93–112.
Fukuyama, H. (2003). Scale characterizations in a DEA directional technology distance function framework. *European Journal of Operational Research, 144*, 108–127.
Golany, B., & Yu, G. (1997). Estimating returns to scale in DEA. *European Journal of Operational Research, 103*, 28–37.
Henriques, I. C., Sobreiro, V. A., Kimura, H., & Mariano, E. B. (2018). Efficiency in the Brazilian banking system using data envelopment analysis. *Future Business Journal, 4*(2), 157–178.
Hung, S. W., Lu, W. M., & Wang, T. P. (2010). Benchmarking the operating efficiency of Asia container ports. *European Journal of Operational Research, 203*, 706–713.
Kleine, A., Dellnitz, A., & Rödder, W. (2014). Sensitivity analysis of BCC efficiency in DEA with application to european health services. In *Operations Research Proceedings* (pp. 243–248).
Kleine, A., Rödder, W., & Dellnitz, A. (2016). Returns To Scale Revisited – towards Cross-RTS. In H Ahn, M Clermont, & R Souren (Eds.), *Nachhaltiges entscheiden: Beiträge zum multiperspektivischen performancemanagement von wertschöpfungsprozessen* (pp. 385–404). Springer-Verlag.
Kounetas, K., Mpourtos, I., & Tsekouras, K. (2009). Efficiency decompositions for heterogeneous technologies. *European Journal of Operational Research, 199*, 209–218.
Lee, C. Y. (2015). Most productive scale size versus demand fulfillment: A solution to the capacity dilemma. *European Journal of Operational Research, 248*, 954–962.
Lovell, C. A. K., & Pastor, J. T. (1995). Units invariant and translation invariant DEA models. *Operations Research Letters, 18*(3), 147–151.
Marshall, A. (1885). On the graphic method of statistics. *Journal of the Statistical Society of London*, 251–260.
Podinovski, V. V. (2004a). Efficiency and global scale characteristics on the "no free lunch" assumption only. *Journal of Productivity Analysis, 22*(3), 227–257.
Podinovski, V. V. (2004b). Local and global returns to scale in performance measurement. *Journal of the Operational Research Society, 55*(2), 170–178.
Podinovski, V. V., Førsund, F. R., & Krivonozhko, V. E. (2009). A simple derivation of scale elasticity in data envelopment analysis. *European Journal of Operational Research, 197*, 149–153.
Ray, S. C. (2007). Are some Indian banks too large? An examination of size efficiency in Indian banking. *Journal of Productivity Analysis, 27*, 41–56.
Ray, S. C. (2015). Nonparametric measures of scale economies and capacity utilization: An application to US manufacturing. *European Journal of Operational Research, 245*, 602–611.
Ren, T. T., Zhou, Z. B., Li, R. Y., & Liu, W. B. (2021). Directional scale elasticity considering the management preference of decision-makers. *RAIRO-Operations Research, 55*, 2861–2881.
Rödder, W., Dellnitz, A., & Litzinger, S. (2022). Combining efficiency and scaling effects in activity analysis: Towards an improved best practice criterion. *RAIRO-Operations Research, 56*, 795–812.
Rödder, W., Kleine, A., & Dellnitz, A. (2017). Scaling Production and Improving Efficiency in DEA: An interactive approach. *Journal of Industrial Engineering International*. https://doi.org/10.1007/s40092-017-0233-7:1-10
Tone, T. (2001). On returns to scale under weight restrictions in data envelopment analysis. *Journal of Productivity Analysis, 16*, 31–47.
Zarepisheh, M., Khorram, E., & Jahanshahloo, G. R. (2010). Returns to scale in multiplicative models in data envelopment analysis. *Annals of Operations Research, 173*, 195–206.
Zhu, J., & Shen, Z. H. (1995). A discussion of testing DMUs' returns to scale. *European Journal of Operational Research, 81*, 590–596.