

Analyzing the accuracy of variable returns to scale data envelopment analysis models

Mansour Zarrin, Jens O. Brunner

Angaben zur Veröffentlichung / Publication details:

Zarrin, Mansour, and Jens O. Brunner. 2023. "Analyzing the accuracy of variable returns to scale data envelopment analysis models." *European Journal of Operational Research* 308 (3): 1286–1301.
<https://doi.org/10.1016/j.ejor.2022.12.015>.



Decision Support

Analyzing the accuracy of variable returns to scale data envelopment analysis models

Mansour Zarrin^a, Jens O. Brunner^{a,b,*}^a Chair of Health Care Operations/Health Information Management, Faculty of Business and Economics, Faculty of Medicine, University of Augsburg, Germany^b Department of Technology, Management, and Economics, Technical University of Denmark, Denmark

ARTICLE INFO

Article history:

Received 9 March 2022

Accepted 12 December 2022

Available online 17 December 2022

Keywords:

Data envelopment analysis

Assurance region

Slacks-based measurement

Variable returns to scale

Monte Carlo data generation

ABSTRACT

The data envelopment analysis (DEA) model is extensively used to estimate efficiency, but no study has determined the DEA model that delivers the most precise estimates. To address this issue, we advance the Monte Carlo simulation-based data generation process proposed by Kohl and Brunner (2020). The developed process generates an artificial dataset using the Translog production function (instead of the commonly used Cobb Douglas) to construct well-behaved scenarios under variable returns to scale (VRS). Using different VRS DEA models, we compute DEA efficiency scores with artificially generated decision-making units (DMUs). We employ five performance indicators followed by a benchmark value and ranking as well as statistical hypothesis tests to evaluate the quality of the efficiency estimates. The procedure allows us to determine which parameters negatively or positively influence the quality of the DEA estimates. It also enables us to identify which DEA model performs the most efficiently over a wide range of scenarios. In contrast to the widely applied BCC (Banker–Charnes–Cooper) model, we find that the Assurance Region (AR) and Slacks-Based Measurement (SBM) DEA models perform better. Thus, we endorse the use of AR and SBM models for DEA applications under the VRS regime.

© 2022 The Author(s). Published by Elsevier B.V.

This is an open access article under the CC BY license (<http://creativecommons.org/licenses/by/4.0/>)

1. Introduction

In order to save resources and to detect inefficient performers, efficiency evaluations are the central component of decision-making management. There are two main classes of efficiency analysis methods in the literature: parametric and non-parametric. Parametric approaches usually use the econometric ordinary least squares method, which shifts regression towards more efficient units to estimate the efficient frontier. This approach is primarily hampered by the assumption about the form of the production function.¹ Contrary to this, non-parametric methods measure efficiency as the distance to an empirical frontier function whose shape is determined by the most efficient decision-making units (DMUs) of the observed dataset. This approach is, without a doubt,

best represented by data envelopment analysis (DEA) introduced by Charnes, Cooper and Rhodes (1978). This model is known as the CCR (Charnes, Cooper, and Rhodes) DEA model. Since the CCR's introduction, a substantial amount of research has been conducted on various aspects of the theory and applications of DEA models. One of these aspects is the economic concept of returns to scale (RTS). There has been much emphasis on the importance of returns-to-scale settings in DEA literature (Dellnitz, Kleine & Rödder, 2018). In this framework, the BCC (Banker, Charnes, and Cooper) DEA model, introduced by Banker, Charnes and Cooper (1984), is the first to assume variable returns to scale (VRS), rather than the CCR's constant returns to scale (CRS). In the literature, both CRS and VRS forms have been developed for almost all upcoming DEA models. Despite this considerable progress over the last five decades, there is still no superior DEA method. Basic models (CCR and BCC) still dominate in various applications, such as healthcare (Kohl, Schoenfelder, Fügenger & Brunner, 2019), despite known concerns including slacks and zero weights. Nevertheless, the development of a *gold standard* can hardly be achieved without a reasonable benchmark with which to compare different DEA models. Due to this lack of operational relevance, DEA is often seen primarily as a scientific topic instead of an operational tool.

* Corresponding author.

E-mail addresses: mansour.zarrin@gmail.com (M. Zarrin), jens.brunner@uni-a.de, jotbr@dtu.dk (J.O. Brunner).¹ An equation that describes the relationship between the number of productive factors (e.g., labor and capital) consumed and the number of outputs produced $D(x, y)$. Production functions can also be used to calculate technical efficiency measures. Suppose that x is used to produce y . The DMU has reached its maximum level of production if $D(x, y) = 1$, given its current level of resources used.

The lack of robustness in results and ambiguity regarding the precision of DEA models' estimates are deemed to be the major quality-related issues. Within the DEA literature, the accuracy and quality analysis of different DEA models have become an attractive area of research over the last two decades. To evaluate the quality of DEA estimates, the first challenge is the absence of *true efficiency* values. DEA estimates in real applications therefore cannot be investigated without these values. Researchers have applied Monte Carlo simulations to create artificial datasets based on certain assumptions and regimes (Cordero et al., 2015) to address this issue. A random distribution function cannot be directly used to derive the scale effect values to reflect the VRS property, so generating well-behaved data is a complicated task. In the following, we summarize the studies conducted on the assessment of the quality of DEA models using Monte Carlo simulations over the last two decades in the interest of brevity. We also discuss the main characteristics of these studies, including the production function used, the number of scenarios, the number of replications, inputs, and outputs. Cobb-Douglas (CD) production functions were most employed by previous studies in the Data Generation Process (DGP) (Holland & Lee, 2002; López, Ho & Ruiz-Torres, 2016; Resti, 2000; Ruggiero, 2005; Simar & Wilson, 2002; van Biesebroeck, 2007). The reason for this can be attributed to the complexities of the alternatives imposing microeconomic regularity conditions like monotonicity and convexity. The limitations of CD for imposing the input substitution elasticity of one and fixed-scale economies have been pointed out by several researchers such as Siciliani (2006) and Perelman and Santín (2009). The Translog² production function has emerged as a generalization of the CD that allows the generation of more testable production data.

Most studies only use one adjustment to account for the number of inputs (López et al., 2016; Ruggiero, 2005). Generally, scenario generation has not been given sufficient attention. Most studies only vary three or fewer characteristics of the employed DGP. Next, previous studies have mainly focused on the properties of the basic DEA models, i.e., CCR and BCC, and comparisons between them and (in some cases) parametric methods (Santín & Sicilia, 2017). However, model evaluations other than the basic ones are rather scarce. So far, only about one-third of previous studies have considered alternative DEA models, and none have utilized more than one model (Kohl & Brunner, 2020). Another concern is the robustness of the results obtained in previous studies. Since the DEA estimations rely on randomly generated data, it is unquestionable that each scenario can be replicated. In this context, Krüger (2012) criticizes the low replication rate of many studies, which changes from 5 to 1000. To our knowledge, the study by Kohl and Brunner (2020) represents the only attempt to date to assess the quality of DEA models by developing meaningful production scenarios using Translog production functions in a CRS setting. The authors develop a sophisticated DGP allowing them to hypothesize some general statements regarding parameters that affect the quality of DEA models through defining some performance indicators. Their results show that the Assurance Region (AR) and Slacks Based Measurement (SBM) models outperform the CCR model under the CRS setting. Kohl and Brunner (2020) primarily discuss the CRS, even though the BCC model remains widely used in most DEA applications (Kaffash, Azizi, Huang & Zhu, 2020; Kohl et al., 2019; Mahmoudi, Emrouznejad, Shetab-Boushehri & Hejazi, 2020).

Last but not least, the literature on DEA focuses mostly on operations research, where the DEA is viewed as a non-econometric or non-statistical approach (Banker, Natarajan & Zhang, 2019; Simar & Wilson, 2015). Thus, a DEA model constructed for assessment needs to move beyond simply explaining and predicting data in

the most effective way possible. In the same way that statistical tests validate a statistical model developed to reproduce accurately the underlying data generation process, basic properties of production economics such as economies of scale and convexity, free disposability, the engineering logic of the production structure, the importance of identified peers to industry participants, etc., serve to validate the model (Banker & Natarajan, 2011; Bogetoft & Otto, 2011b). By identifying conditions under which DEA estimators are statistically consistent and likelihood-maximizing, Banker (1993) provided a formal statistical basis for DEA. Accordingly, DEA estimates are capable of providing interesting insights without heavily relying on statistical testing. However, most of the literature ignores the statistical properties of the estimators and lacks consistent statistical tests to compare the efficiencies between two samples. These researchers compare their improvements to the basic model and highlight properties such as a shift in the average efficiency scores or a better discrimination power. Even if a certain problem can be solved through development, there is no guarantee that the overall results (from a quality perspective, for example) will also be improved. The main flaw here is comparing differences in DEA estimations through the mean value of the efficiency scores rather than the distribution of them. However, in cases where the distribution of efficiency scores is skewed, the mean value becomes an ineffective measure of central tendency (Weisberg, 1992). Several studies have been performed on comparing differences in DEA estimation³ distributions for two groups of DMUs through developing statistical tests including parametric and non-parametric ones. For example, Cummins, Weiss and Zi (1999) use a regression-type parametric test with a dummy variable indicating the groups, regressing the efficiency scores on the dummy variable. However, many researchers (e.g., Golany and Storbeck (1999) and Lee, Park and Choi (2009)) believe that non-parametric tests such as the Mann-Whitney and Kruskal-Wallis tests are more appropriate since they do not make assumptions on the distribution of efficiency scores. One pioneering study in this direction has been conducted by Banker, Zheng and Natarajan (2010). They develop two sets of parametric and three non-parametric tests and compare them against the F-tests introduced by Banker (1993). They show that their developed tests outperform the F-tests in Banker (1993) when noise plays an important role in the data generating process. However, the F-tests in Banker (1993) remain effective if efficiency dominates noise. In our study, we integrate the idea of comparing two groups of DMUs with the performance indicators.

The purpose of this study is to address these issues by providing a method for evaluating the accuracy of DEA models under the VRS assumption. A sophisticated DGP must be designed to create well-behaved data for the DMUs to study the quality of DEA models. In the next step, we generate artificial data so that the true efficiency of each DMU can be compared with the estimations obtained from the different DEA models. Through this, we are able to evaluate the DEA models' quality. We then consider a variety of scenarios to arrive at generally sound conclusions. With these characteristics, it is possible to generate meaningful data through Monte Carlo Simulations. We use two aggregated benchmark values: benchmark value (B-Value) and benchmark rank (B-Rank). Combined with multiple performance indicators, these benchmark values cover all relevant properties of an efficiency estimator, such as identifying efficient and inefficient units and ranking the efficiency score of each unit in a set of DMUs. The B-Value and B-Rank provide additional insight into the performance of the procedure by using SBM, AR, the basic CCR DEA,

² Translog stands for transcendental logarithmic.

³ In many studies, the terms "inefficiency" and "efficiency" are interchangeably used with each other to describe the scores obtained by DEA models.

and uniformly distributed random numbers (Rand). Based on our findings, we conclude that the environment of a DEA application influences its results significantly. We do this by casting doubt on the reliability of DEA results and analyzing the efficiency assessment process of the DEA model. We analyze the VRS settings as the most prevalent setting in the literature for DEA applications and try to find out whether the predominant BCC position is justified. Our study addresses the statistical properties of DEAs' estimators by applying a consistent statistical test to compare the estimations calculated based on different DEA models with the true efficiencies. The details of our analysis will be presented in subsequent sections. As a summary, this paper contributes the following to the pertinent literature:

- I. The main question this study seeks to answer is whether BCC's dominant position was indeed vindicated. To do this, we analyze and compare the BCC model estimates with two other DEA models: AR and SBM. Comparisons with the basic model for BCC DEA and uniformly distributed random numbers (i.e., Rand) reveal also the accuracy of the procedure.
- II. Two approaches are used to conduct the comparison: benchmark scores based on multiple performance indicators and DEA-based hypothesis tests. Benchmark scores cover many aspects of a measure of efficiency introduced by Pedraja-Chaparro, Salinas-Jiménez and Smith (1999), such as identifying the most efficient DMUs and ordering their efficiency scores within a sample. We acknowledge the need for a statistical foundation for DEA as pointed out by Banker (1993), Banker et al. (2010), and Simar and Wilson (2015), and test the estimations of DEA models with their actual efficiencies by running statistical tests.
- III. In order to improve the general validity of our results, we advance the scenario variation significantly. In our study, each generated scenario represents an arrangement of varying values for different characteristics of the DGP (e.g., number of inputs, number of DMUs, the importance of input). With 7776 scenarios generated based on the VRS setting, we attain the highest level of validity in the quality assessment of VRS DEA models in comparison to the literature. To determine whether the environment of the DEA study influences the accuracy of results, we also consider the coverage of different characteristics. By utilizing ten different characteristics with varying levels, we provide another significant contribution to the literature.
- IV. The general form of Translogs has the consequence of not being monotonic or globally convex like CDs. For generating well-behaved data under the VRS setting, we need to impose the necessary curvature requirements on a Translog, which is a challenging problem (Greene, 2008). Then we propose a mathematical model that directly enforces monotonicity and curvature requirements and generates valid scenarios with VRS properties. Using our methodology, one can modify the input substitution in order to ensure a more sensible DGP. According to the literature, a handful of studies, like Krüger (2012), consider different input substitutions using Constant Ratio of Elasticity of Substitution Homothetic or Constant Elasticity of Substitution production functions. Through several adjustable parameters, the Translog production function offers greater control over setting input substitutions. Setting these parameters to generate valid scenarios (or well-behaved data), however, is a complicated process. As a result, only a few studies use it in a limited form to generate the data. For example, Cordero et al. (2015), who focus on generating data under decreasing returns to scale (DRS), or Perelman and Santín (2009), who define the parameters arbitrarily. We advance the approach used by Kohl and Brunner (2020) for the CRS setting so that realistic scenarios under the VRS regime can be generated systematically.

- V. By decomposing the input substitution into two terms: substitutability and distribution of substitutions, we are able to guarantee the generation of realistic and well-behaved DMUs under the VRS, along with a variety of scenarios. We find a high correlation between the number of replications for each scenario and the number of DMUs from the perspective of the robustness of the results. A scenario with 450 DMUs may need 50 replications while a small size scenario (e.g., 50 DMUs) might need over 200 replications. We, therefore, define an elastic stopping condition for replications of each scenario based on the moving standard deviation (StD) of the benchmark value. Finally, we examine the impact of the characteristics considered in the generation of the distinctive scenarios (e.g., sample size) on the quality of estimations calculated using the different DEA models.

The rest of this study is structured as follows. Section 2 describes in detail the steps of developing a DGP, statistical tests, performance indicators, and study design. In Section 3, the results of comparisons are presented and discussed in detail. Finally, the paper is concluded in Section 4.

2. Methodology

We describe all steps within the proposed framework thoroughly in the following subsections, in order to compare and analyze the accuracy of DEA models within a VRS context. Fig. 1 depicts the eight steps of the DGP for every DMU.

2.1. Performance indicators

Following the purpose of evaluation and comparison of different DEA models, we utilize five performance indicators defined by Kohl and Brunner (2020) (see Appendix A) based on Pedraja-Chaparro et al. (1999) for Monte Carlo DEA analyses. The DEA's estimates are the core of any judgment on the quality. Therefore, for defining the performance indicators, we address the four main purposes of a DEA containing recognizing inefficient DMUs, ranking the efficiency of DMUs, assessing efficiencies and rooms for improvement, and investigating the overall efficiency of a company/organization.

2.2. Hypothesis tests for comparing efficiency

We compare the efficiency distribution of two groups of DMUs using DEA-based hypothesis tests in addition to the performance indicators. Constructing statistical tests allows us to evaluate the null hypothesis of no difference in the distributions of true efficiency (θ) and estimated efficiency ($\hat{\theta}$) obtained from the DEA models. The null hypothesis of no difference in efficiency distributions of true efficiency can be tested using the procedure proposed by Banker (1993). The first step of this method is to determine whether the efficiency scores are normally or exponentially distributed. The true efficiency in our DGP is normally distributed. Now suppose both θ and $\hat{\theta}$ are distributed as normal with parameters ρ_1 and ρ_2 , respectively. Then, the test statistic can be calculated as $(\sum_j (\theta_j)^2/n)/(\sum_j (\hat{\theta}_j)^2/n)$ under the null hypothesis of no difference between them (i.e., $H_0 : \rho_1 = \rho_2$), and compared with the critical value of the F distribution with (n, n) degrees of freedom at the significance level of 5%. Banker et al. (2010) evaluate the performance of this test against the other parametric (e.g., T-test) and non-parametric (e.g., Mann-Whitney's U test) tests used traditionally in the DEA literature (Banker & Natarajan, 2011). Their

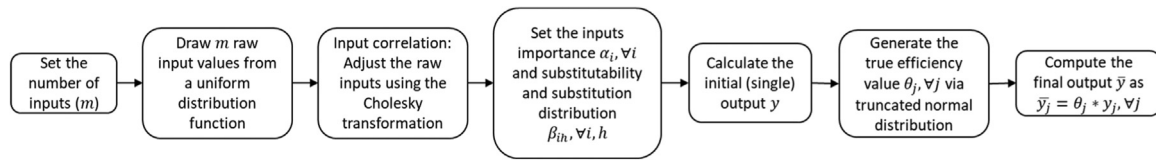


Fig. 1. Developed DGP for each artificial DUM.

simulation results indicate this test is adequate for detecting deviations from the efficiency frontier caused by a single inefficiency term.

2.3. Data generation process under VRS setting

This paper extends the sophisticated DGP proposed by Kohl and Brunner (2020) for the CRS setting to generate well-behaved production data with the VRS system. The DGP produces a single output (y) based on the generated meaningful inputs (x_i, i ∈ M = {1, ..., m}) and true efficiency values (θ_j) for each DMU in which the regularity conditions are met. According to this information, the technology can be shown by the graph set T = {(x, y) : x can produce y}. The hyperbolic output distance function can be introduced as the maximum equiproportionate expansion of an output vector and reduction of an input vector that places an observation within the boundary of a technology T, i.e., D_O(x, y) = inf {θ : θ > 0, (x, y/θ) ∈ T}, if the graph production possibility set satisfies the axioms described in Coelli, Prasada Rao, O'Donnell and Battese (2005). We generate the well-behaved dataset by using the (logarithmic) Translog production function presented by Eq. (1). This technology has become the gold standard for Monte Carlo simulations (Bogetoft & Otto, 2011a).

$$\ln D_{Oj}(\mathbf{x}, y) = \alpha_0 + \ln y_j + \sum_{i=1}^m \alpha_i \ln x_{ij} + \frac{1}{2} \sum_{i=1}^m \sum_{h=1}^m \beta_{ih} \ln x_{ij} \ln x_{hj}, \quad \forall j = 1, \dots, n \quad (1)$$

where, α₀ is the efficiency parameter (can be set as 0), y is the initial output, parameters α_i and β_{ih} show respectively the importance of an input i, and the substitution possessions of the production procedure between two inputs i and h. These parameters are defined to acquire a well-behaved production function within the boundaries imposed by the inputs (x_i). We develop a seven-step DGP for each DMU under the VRS setting (depicted in Fig. 1) by ensuring adherence to the properties defined by Coelli et al. (2005) for well-behaved VRS data. In our DGP, apart from generating the parameters α and β, true efficiency (θ), input vector x (including the number of inputs (m), input range, and input correlation), and the regularity conditions (monotonicity, curvature, and quasi-convexity) are meticulously taken into consideration to generate valid scenarios.

The value of true efficiency (θ) is drawn from a truncated normal distribution and then multiplied by the raw output value. We include different true efficiency distributions in our DGP as an adjustable characteristic to examine whether the true efficiency level influences the accuracy of VRS DEA models. The truncation is always set at 1.0 for the upper-efficiency values. Different lower bounds can be set to imitate diverse economies of scale. By adjusting the mode and standard deviation (StD) of the true efficiencies, a comparable distribution shape can be preserved. We then calculate the final output \bar{y} by multiplying the initial output by the true efficiency value: $\bar{y}_j = \theta_j \cdot y_j$.

Adjusting the number of inputs, the range of inputs, and the correlation among inputs all lead to the generation of the input vector x. Adjustments are generally straightforward, for example,

changing the number of inputs and parameters of the uniform distribution function used for the level of inputs. The wide range of inputs indicates a more heterogeneous production environment. Instead, the small range of inputs suggests a very homogeneous dataset with entities of similar sizes. A correlation between the input values also seems logical as larger entities usually use more inputs than smaller ones. A Cholesky decomposition method described in Hazewinkel (1992) accounts for this fact when generating inputs.

An authentic VRS production data requires the change of scale effects with the size of the DMU. Therefore, an optimal size must be defined within the economically feasible region⁴ of production, at which the average product is maximized. For example, in the case of a single-input single-output production function, the average product is y_1/x_1 where graphically represents the slope of the line (ray) that passes through the origin and that point. This point is known as the point of optimal scale (of operations) where units exhibit CRS, smaller units work under increasing returns to scale (IRS) and bigger ones work under the DRS setting (Coelli, Rao & Battese, 1998). We represent units that have exactly the optimal scale of operations as \mathbf{x}^{CRS} . Then, the necessary conditions of VRS setting for returns to scale can be written as Eq. (2) by straightforward operations on Eq. (1) (Balk, 2001). The scale elasticity value of DMU_j of the output distance function defined in Eq. (1) (Balk, 2001) is:

$$\phi_{Oj}(\mathbf{x}_j, y_j) = \sum_{i \in M} \frac{\partial \ln y_j}{\partial \ln x_{ij}} = \sum_{i \in M} \alpha_i + \sum_{i \in M} \left(\beta_{ii} + \sum_{h \in M \setminus \{i\}} \beta_{ih} \right) \ln x_{ij} \quad (2)$$

$$\ln x_{ij} \begin{cases} > 1 \leftrightarrow IRS \\ \frac{1}{2} < 1 \leftrightarrow CRS \\ < 1 \leftrightarrow DRS \end{cases}, \quad \forall j = 1, \dots, n$$

where φ_{Oj}(x_j, y_j) represents the (output distance function based) scale elasticity value of DMU_j at point (x_j, y_j). Note that the symbol “!” above the equal and unequal signs means “must hold”. If this value is greater than, equal to, and lower than 1, we respectively have IRS, CRS, and DRS.⁵ The scale elasticity in Eq. (2) is decomposed into two terms $\sum_{i \in M} \alpha_i$ and $\sum_{i \in M} (\beta_{ii} + \sum_{h \in M \setminus \{i\}} \beta_{ih}) \ln x_{ij}$. The first term represents the importance of inputs and the second one sets input substitutability and substitution distribution. According to these two terms, we can define the sufficient conditions for satisfying the global VRS regime that still allows the implementation of substitution effects for each DMU_j as: $\sum_{i \in M} \alpha_i > 1 \cap \sum_{i \in M} (\beta_{ii} + \sum_{h \in M \setminus \{i\}} \beta_{ih}) \ln x_{ij} < 0$ (∩ means AND).

For the data generation process, we want to test different optimal sizes as well as the extent of the economics scale effects. For that reason, we can reformulate $\sum_{i \in M} \alpha_i > 1$ as $\sum_{i \in M} \alpha_i = 1 + \omega$, ω > 0 which satisfies the first sufficient condition we need to

⁴ A region where is consistent with all properties defined for the production function such as monotonicity.

⁵ The corresponding output scale efficiency value SE_{Oj}(x_j, y_j) for DMU_j can be calculated by $\ln SE_{Oj}(x_j, y_j) = -(D_0(x_j, y_j) - 1)^2 / 2 \sum_{i=1}^m \sum_{h=1}^m \beta_{ih}$, $\forall j$ as indicated by Balk (2001).

guarantee the global VRS regime. The parameter ω can be used to adjust the extent of scale effects. A small ω implies weak scale effects, while the revert is true for a large value. We can implement different adjustments for the input importance by altering the value of α . We here apply two different adjustments containing equal and equidistant importance. In both settings, we must hold $\sum_{i \in M} \alpha_i = 1 + \omega$, $\omega > 0$ to guarantee the implementation of the VRS regime. In the first adjustment (hereafter referred to as SYM), every input is identically important in the production function. This can be achieved by Eq. (3). The definition provided in Eq. (3) for α_i fulfills the condition of $\sum_{i \in M} \alpha_i > 1$. It is proven in Appendix B (Proposition 1).

$$\alpha_i = \frac{1 + \omega}{m}, \quad \forall i \tag{3}$$

The second setting (hereafter referred to as ASYM) generates a production function with inputs of varying importance yet equidistant (see Eq. (4)). In this adjustment, the first input (x_1) is always the one with the lowest influence on production, and the importance of the other inputs increases with their indices. Consider three inputs x_1 , x_2 , and x_3 , since x_1 has the smallest importance (smallest index) to the production process, one unit increase in it would lead to a lesser rise in the output level than one unit increase in either x_2 or x_3 does. Of these, x_3 would lead to the largest growth in output. Since we only consider abstract inputs that can be rearranged, there will be no misrepresentation of the results due to this regularity. The definition provided in Eq. (4) fulfills the condition of the VRS setting (i.e., $\sum_{i \in M} \alpha_i > 1$) as proven in Appendix B (Proposition 2).

$$\alpha_i = \frac{(1 + \omega) \cdot (i + m)}{1.5m^2 + 0.5m}, \quad \forall i \tag{4}$$

The second term of Eq. (2) i.e., $\sum_{i \in M} (\beta_{ii} + \sum_{h \in M \setminus \{i\}} \beta_{ih}) \ln x_i$, which deals with β parameters should be less than or equal to zero to ensure the VRS regime. β represents the substitution of two inputs and must satisfy the symmetry condition $\beta_{ih} = \beta_{hi}$, $\forall i, h$ (Coelli et al., 1998). Note that the condition of linear homogeneity of degree +1 in outputs is automatically satisfied in a single-output case (Coelli et al., 1998). Having in mind $\sum_{i \in M} \alpha_i = 1 + \omega$, the second term of Eq. (2) must be exactly equal to $-\omega$, in other words, $\sum_{i \in M} (\beta_{ii} + \sum_{h \in M \setminus \{i\}} \beta_{ih}) \ln x_i = -\omega$ to achieve CRS at x^{CRS} , i.e., the optimum technical efficient size. This property can be fulfilled by Eq. (5) where it is assumed that the optimum technical efficient size of all inputs is at the same point, x^{CRS} (i.e., $x_i^{CRS} = x^{CRS}$, $\forall i$).

$$\sum_{i \in M} \left(\beta_{ii} + \sum_{h \in M \setminus \{i\}} \beta_{ih} \right) = -\frac{\omega}{\ln x_i^{CRS}} \tag{5}$$

The β parameters are responsible for satisfying two main economic regularity properties: monotonicity (or non-decreasing) and concavity (or non-increasing) in all inputs (Coelli et al., 2005). Taking into account these properties, β cannot be set freely. We decompose β into two terms: substitution distribution (σ_{ih}) and substitutability (ν), mathematically, $\beta_{ih} \propto \sigma_{ih} \cdot \nu$, $\forall i, h$. This decomposition advantages us in adjusting both characteristics substitutability and substitution distribution separately in our DGP as well as in examining their possible effects on the accuracy of DEA estimates. The substitution distribution (σ_{ih}) deals with the fact that the inputs substitution might be identical between all inputs and it is responsible for the distribution of β . The substitutability (ν) characteristic determines the magnitude of β to be able to consider fluctuating capabilities to substitute inputs. Since the final magnitude of β should be regulated by its substitutability (ν), the substitution distribution (σ_{ih}) are normalized between -1 and 1 . Referring to the symmetry condition, it must hold $\sigma_{ih} = \sigma_{hi}$, $\forall i, h$.

We can reflect the possible effects of the substitution distribution (σ_{ih}) by defining two different settings: *equal* where the substitution between all inputs is equal (Eq. (6)); and *unequal* where we advance the pattern proposed by Kohl and Brunner (2020) to generate unequal yet symmetric values for β_{ih} , $\forall i, h$. In both equal and unequal settings, we need to satisfy the condition presented by Eq. (5) as well as the symmetry to guarantee the implementation of the VRS setting through the substitution distribution (σ_{ih}).

For the equal substitution distribution, we have $\sum_{i \in M} (\beta_{ii} + \sum_{h \in M \setminus \{i\}} \beta_{ih}) = m \cdot (\beta_{ii} + \sum_{h \in M \setminus \{i\}} \beta_{ih})$ by construction, as a result, we can rewrite Eq. (5) as $\beta_{ii} + \sum_{h \in M \setminus \{i\}} \beta_{ih} = -\frac{\omega}{m \cdot \ln x_i^{CRS}}$, $\forall i$. Definitions provided in Eq. (6) respecting Eq. (5) are proven in Appendix B (Proposition 3).

$$\beta_{ii} = \frac{-\nu \cdot \omega}{m \cdot \ln x_i^{CRS}}, \quad \forall i \text{ and } \beta_{ih} = \frac{(\nu - 1) \cdot \omega}{m \cdot (m - 1) \cdot \ln x_i^{CRS}}, \quad \forall \{i, h | i \neq h\} \tag{6}$$

Imposing the equal or identical substitution distribution is simple and can be accomplished by defining $\sigma_{ii} = -\frac{1}{m}$, $\forall i$ and $\sigma_{ih} = \frac{1}{m(m-1)}$, $\forall \{i, h | i \neq h\}$. Therefore, we can rewrite the definitions of β provided in Eq. (6) as follows:

$$\beta_{ii} = \frac{\nu \cdot \omega}{\ln x_i^{CRS}} \cdot \sigma_{ii}, \quad \forall i \text{ and } \beta_{ih} = \frac{(\nu - 1) \cdot \omega}{\ln x_i^{CRS}} \cdot \sigma_{ih}, \quad \forall \{i, h | i \neq h\} \tag{7}$$

For modeling the unequal substitution scenario, we develop the pattern presented by Kohl and Brunner (2020), to create symmetric but unequal values for β via formulas presented in Eq. (8) with $\sigma'_{ii} = -\frac{m \cdot (1.5 - \frac{i-1}{m-1}) - (2 - 2 \cdot \frac{i-1}{m-1})}{1.5m-2}$, $\forall i$ and $\sigma'_{ih} = \frac{2 - \frac{h-1}{m-1} - \frac{i-1}{m-1}}{1.5m-2}$, $\forall \{i, h | i \neq h\}$ given in Kohl and Brunner (2020). These definitions also respect Eq. (5) as shown in Appendix B (Proposition 4).

$$\beta_{ii} = -\frac{\omega \cdot (1 - \nu \cdot \sigma'_{ii})}{m \cdot \ln x_i^{CRS}}, \quad \forall i \text{ and } \beta_{ih} = \frac{\omega \cdot \nu}{m \cdot \ln x_i^{CRS}} \cdot \sigma'_{ih}, \quad \forall \{i, h | i \neq h\} \tag{8}$$

Now, we turn to the substitutability of inputs controlled by parameter ν . Substitutability boundaries differ for certain inputs. Again, the monotonicity of the production function is the source of the substitutability conditions. For single-output multi-input, monotonicity implies constraints on partial derivatives of distance functions. These constraints can be expressed by Eq. (9). The mandatory curvature and monotonicity conditions of the production function are key factors in the characteristics of well-behaved production data (Cordero et al., 2015; Perelman & Santín, 2009). The partial derivatives of distance functions must satisfy one condition for monotony: for D_0 as a single output, all marginal products (f_i) must be non-negative across all inputs (x_i) as outlined by Eq. (10).

$$s_i = \frac{\partial \ln D_0}{\partial \ln x_i} = \alpha_i + \sum_h \beta_{ih} \ln x_h, \quad \forall i \tag{9}$$

$$f_i = \frac{\partial D_0}{\partial x_i} = \frac{\partial \ln D_0}{\partial \ln x_i} \frac{D_0}{x_i} = s_i \frac{D_0}{x_i} \geq 0 \Leftrightarrow s_i \geq 0, \quad \forall i \tag{10}$$

Curvature guarantees that all marginal products must be declining, i.e., the law of diminishing marginal productivity (Coelli et al., 2005). The condition can be satisfied by fulfilling Eq. (11) which is the second partial derivative obtained by applying the chain rule to Eq. (1).

$$f_{ii} = \frac{\partial^2 D_0}{\partial x_i \partial x_i} = \frac{\partial f_i}{\partial x_i^2} = \frac{\partial (s_i \frac{D_0}{x_i})}{\partial x_i} = (\beta_{ih} + s_i s_i - s_i) \left(\frac{D_0}{x_i^2} \right) < 0 \Leftrightarrow \beta_{ih} + s_i s_i - s_i < 0, \quad \forall i \tag{11}$$

For quasi-convexity in inputs, the corresponding bordered Hessian matrix $F(x_i)$ (Eq. (12)) on inputs needs to be evaluated.

$$F(x_i) = \begin{bmatrix} 0 & f_1 & f_2 & \cdots & f_i \\ f_1 & f_{11} & f_{12} & \cdots & f_{1i} \\ f_2 & f_{21} & f_{22} & \cdots & f_{2i} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ f_i & f_{i1} & f_{i2} & \cdots & f_{ii} \end{bmatrix} \quad (12)$$

where, $f_{ih} = \frac{\partial^2 D_0}{\partial x_i \partial x_h} = \frac{\partial f_i}{\partial x_i \partial x_h} = \frac{\partial (s_i \frac{D_0}{x_i})}{\partial x_h} = (\beta_{ih} + s_i s_h) (\frac{D_0}{x_i x_h})$, $\forall \{i, h | i \neq h\}$, f_i and f_{ii} have been already defined by Eqs. (10) and (11), respectively. The isoquants are strictly quasi-convex on inputs if this bordered Hessian matrix is negative definite (Coelli et al., 2005). $F(x_i)$ is negative definite if the successive principle minors alternate in sign. Defining the $i + 1$ principle minor by $F(x_i)$, F is negative definite if $(-1)^i |F^i(x)| > 0$.

The expressive DGP should ensure that an increase in inputs does not lead to a decline in output despite changing the substitutability of inputs. It echoes the concept of input-free disposability found in the vast majority of DEA models. Keeping the curvature and monotonicity constraints is critically dependent on the magnitude of β . Therefore, we present the mathematical programming approach as Model (13) to derive the optimum value of ν that allows modifying the substitutability between inputs. Having a minimum value of ν gives a nearly flat substitution curve, resulting in high substitutability, while a maximum value of ν results in low substitutability.

$$\min / \max \nu \quad (13a)$$

$$s.t. \ s_i \geq 0, \ \forall i \quad (13b)$$

$$\beta_{ih} + s_i^2 - s_i < 0, \ \forall i \quad (13c)$$

$$(-1)^i |F^i(x)| > 0, \ \forall i \quad (13d)$$

Values of the first and second partial derivatives, i.e., s_i and f_{ii} , fluctuate with input levels then, we cannot generally guarantee that the isoquants are strictly convex (Coelli et al., 1998). However, as explained by Coelli et al. (1998), there are areas in the input space where Eqs. (10) and (11) are satisfied. Providing that these conditions can be satisfied for every data point for any proposed Translog function, the well-behaved area may be large enough to adequately represent the corresponding production function. Note that the constraints of Model (13) change according to the number of inputs as the bordered Hessian matrix changes. The curvature and quasi-convexity inequalities (Eqs. (13c) and (13d)) are quadratic and nonlinear, respectively. These constraints make solving the optimization problem considerably more difficult. In the two-input single-output case ($i = 1, 2$), the model and the bordered Hessian matrix in the quasi-convexity (the third constraint, i.e., (13d)) can be rewritten by considering the definitions of β_{ii} and β_{ih} provided in Eq. (6), as follows:

$$\min / \max \nu \quad (14a)$$

$$s.t. \ s_1 = \alpha_1 + \beta_{11} \ln x_1 + \beta_{12} \ln x_2 \geq 0 \quad (14b)$$

$$s_2 = \alpha_2 + \beta_{22} \ln x_2 + \beta_{21} \ln x_1 \geq 0 \quad (14c)$$

$$f_{11} = \beta_{11} + s_1^2 - s_1 < 0 \quad (14d)$$

$$f_{22} = \beta_{22} + s_2^2 - s_2 < 0 \quad (14e)$$

Table 1
Defined characteristics for generating scenarios.

Characteristic	Value/Level
Returns to scale	VRS
True efficiencies (θ)	Low, Medium, High
# DMUs (n)	50, 150, 450
# Inputs (m)	2, 5, 7
Importance of inputs (α_i)	SYM and ASYM
Input substitutability (ν)	Low and High
Input substitution distribution (β_{ih})	Equal and Unequal
Input range	U[100; 1100] and U[100; 10,100]
Input correlation	0.0, 0.4, 0.8
Efficient size (x_i^{CRS})	300, 600
Extent of scale effects (ω)	0.2, 0.4, 0.8
Total Number of Scenarios	7776

$$(-1)^1 |F^1| > 0 \Leftrightarrow |F^1| = 0 * f_{11} - f_1 * f_1 = -f_1^2 < 0 \quad (14f)$$

$$(-1)^2 |F^2| > 0 \Leftrightarrow |F^2| = f_1 f_{12} f_2 - f_1 f_1 f_{22} + f_2 f_1 f_{21} - f_2 f_{11} f_2 > 0 \quad (14g)$$

We reformulate the model to transform the nonlinear constraints into a minimal number of conjunctive linear constraints that have the same admissible marking area as the nonlinear one does. The first quasi-convexity condition (Eq. (14f)) is fulfilled since the first principal minor $|F^1|$, is always negative. For $i = 2$, the second principal minor $|F^2|$ (Eq. (14g)), can be written as $2f_1 f_2 f_{12} - f_1^2 f_{22} - f_2^2 f_{11}$. This expression should be positive to guarantee the necessary and sufficient condition of quasi-convexity in inputs. The term $-f_1^2 f_{22} - f_2^2 f_{11}$, which is equivalent to $-s_1^2 \frac{D_0^3}{x_1^2 x_2^2} (\beta_{22} + s_2^2 - s_2) - s_2^2 \frac{D_0^3}{x_1^2 x_2^2} (\beta_{11} + s_1^2 - s_1)$, is always positive by construction. Consequently, we can simply show that one sufficient condition to fulfill Eq. (14g) is that the term $f_1 f_2 f_{12}$ be non-negative. From Eqs. (14b) and (14c), we know that f_1 and f_2 are non-negative. Therefore, one sufficient condition to assure quasi-convexity is:

$$f_{12} = (\beta_{12} + s_1 s_2) (\frac{D_0}{x_1 x_2}) \geq 0 \Leftrightarrow \beta_{12} \geq 0 \quad (15)$$

The impositions of the α and β values play the main role in the design of scale elasticity as well as in the computation of scale efficiency scores. A well-behaved production function can be obtained with the proposed model by imposing desirable assumptions. There is no doubt that increasing the number of inputs also increases the number of regularity conditions to which the proposed mathematical model must submit. Nevertheless, the procedures described for the two-input sample can be adapted to cases with higher multi-input dimensions. By sizing up the dimension of the problem, the proposed model can be used to generate regular behaved data, which would otherwise become cumbersome. Now that all the characteristics are adjustable, a well-behaved DMU can be generated under the VRS setting.

2.4. Study design

The characteristics used in this study are listed in Table 1 along with their values/levels. After creating one scenario as an example, the obtaining dataset is assessed using four different output-oriented DEA models: CCR (Charnes et al., 1978), BCC (Banker et al., 1984), VRS AR⁶ (Pedraja-Chaparro, Salinas-Jimenez & Smith,

⁶ Since we deal with different input elasticities, we apply virtual weight restrictions (the product of weight and input/output) in the AR model. We set k to limit the virtual weights to 2 as Pedraja-Chaparro et al. (1997) did.

Table 2
An example scenario for the two-input single-output case.

Characteristics	Value/Level
True efficiencies (θ)	Medium
# DMUs (n)	50
# Inputs (m)	2
Input range	$U[100; 1100]$
Input correlation	0
Efficient size (x_i^{CRS})	300
Extent of scale effects (ω)	0.2
Importance of inputs (α_i)	SYM
Input substitutability (ν)	Low
Input substitution distribution (β_{ih})	Equal

1997), and VRS SBM (Tone, 2001). The DEA models under study are described in Appendix C. Moreover, we compute the benchmark model Rand, which consists of randomly drawn values similar to the real efficiency distribution, to ensure a thorough comparison of VRS DEA models with Monte Carlo simulated data. In theory, Rand provides a lower bound for benchmark values and allows the classification of B-Values derived from DEA models. DEA applications fall into three categories according to the number of DMUs: small (50 DMUs), medium (150 DMUs), and large (450 DMUs).

The number of DMUs in the generated scenarios can be modified by simply running the DGP for one DMU n times. The true efficiency score θ , as mentioned before, is drawn from the truncated normal distribution and multiplied by the raw output y_j for each DMU. Using true efficiency distributions as characteristics, we examine whether the level of true efficiencies influences the accuracy of DEA models. In the true efficiency score distributions, the upper bound is always set at 1.0, but the lower bound can be customized based on three different values: low (0.25), medium (0.40), and high (0.55). These levels reflect the reality that poor-efficiency DMUs cannot survive. Changing the modes and StD of true efficiencies will result in similar curves. Therefore, we use modes of 0.75 (low), 0.80 (medium), and 0.85 (high) and StDs of 0.27 (low), 0.25 (medium), and 0.23 (high). For each DMU, the value of m inputs is randomly selected from two uniform distributions: $U[100; 1, 100]$ and $U[100; 10, 100]$. The ranges used here have been derived from a study conducted by Kohl and Brunner (2020); they compared various ranges to determine the most meaningful ones. In addition, the Cholesky decomposition is applied to impose the correlation coefficients of 0.0, 0.4, and 0.8 between the raw inputs as described in Hazewinkel (1992).

3. Results and discussions

Our main objective is to evaluate the accuracy of four main DEA models and to determine the scale efficiency of generating scenarios based on the defined characteristics. The results are divided into three parts. First, we intend to make the results more understandable by introducing some numerical illustrations explaining the characteristics used for generating scenarios. Our next task is to present the results of our main computational study. This will enable us to figure out which models of DEA based on the VRS setting perform best and to explore the driving factors. Our final section provides guidelines on how to apply DEA models in VRS settings based on our computational results.

3.1. Numerical illustrations

For the two-input single-output case, we generate the well-behaved production function based on the Translog output distance function described before. Considering the settings given in Table 2, we calculate the values of α_i , ν , β_{ih} using Eq. (3), Model (14), and Eq. (6), respectively. For a given input vector (e.g., $\mathbf{x} =$

Table 3
Results of the two-input single-output instance.

Characteristics	Values
Importance of inputs (α_i)	$\alpha = [0.6, 0.6]$
Input substitutability (ν)	$\nu = 12.3514$
Input substitution distribution (σ_{ih})	$\sigma = \begin{bmatrix} -1 & 1 \\ 1 & -1 \end{bmatrix}$
Input substitution (β_{ih})	$\beta = \begin{bmatrix} -0.2165 & 0.1990 \\ 0.1990 & -0.2165 \end{bmatrix}$
Monotonicity conditions ($s_i \geq 0$)	$s_1 = 0.8758$ and $s_2 = 0.1312$
Curvature conditions ($f_{ii} < 0$)	$f_{11} = -0.3252$ and $f_{22} = -0.3305$
Quasi-convex in inputs ($(-1)^i F^i(\mathbf{x}) > 0$)	$ F^1 = -0.7671$ and $ F^2 = 0.3314$

[100; 1, 100]), the obtained values are presented in Table 3. In Appendix D, we provide the dataset generated for this instance. If we set x_i^{CRS} close to the minimum of our input range (100), the change of scale effects according to the size of the DMU starts at the beginning of the production function. This effect of x_i^{CRS} is shown in Fig. 2(a) in which we represent the production function of 1000 DMUs under two different values of 300 and 600 and the same setting for the other characteristics as reported in Table 2. The effect of ω which is responsible for adjusting the extent of scale effects, for two different values of 0.2 and 0.4 is shown in Fig. 2(b). As the value of ω increases, the curvature of the production function also increases. According to the minimum and maximum of ν , which allow the adjustment of the substitutability, high and low substitutability are recommended between inputs. Fig. 2(c) shows the effect of substitutability on the production function. We see that the minimum value of ν produces almost a level surface without large raised areas or indentations, while the maximum value of it produces a curve-shaped surface.

3.2. Results of analyzing the accuracy of DEA models

In the following sections, we discuss the results of the evaluation of four VRS DEA models and the Rand data gathered from 7776 scenarios. In Table 4, we report the minimum (Min), maximum (Max), mean, and StD values of the performance indicators over all scenarios. In addition, boxplots depict the main descriptive statistics of B-Values and B-Ranks for each model in Fig. 3. We use the Rand model as a lower bound for our benchmarks. The average, maximum, and minimum number of replications required for each scenario are respectively 111, 270, and 50. We define the stopping criterion for the replication based on the moving StD of the B-Value for the DEA models. If the moving StD of the B-Value of all four DEA models is less than 0.001, the replication terminates. There are over 434,000 replications in all, and each replication is tested using all four DEA models. By construction, we impose VRS technology on the DGP so that the efficiency scores calculated with DEA models under the VRS setting should be better than those calculated with CRS DEA models. To compute scale inefficiency as well as evaluate the potential bias associated with computing efficiency scores under CRS when true technology is represented by VRS, we run the CCR model. CCR results emphasize the importance of using an accurate return to scale before conducting a practical DEA efficiency analysis. Consider, for instance, the mean B-Value of the CCR, which is equal to 0.295, and its VRS counterpart (BCC), which is almost double, 0.574.

The small value of Mean Absolute Error (MAE) suggests the estimated efficiency scores are on average close to their true counterparts, and therefore, high $1 - MAE$ values are preferred. According to Table 4, the MAE cannot provide information about the deviation because of the small mean value of this indicator for Rand = 0.824) which is very close to the VRS DEA models. In order to

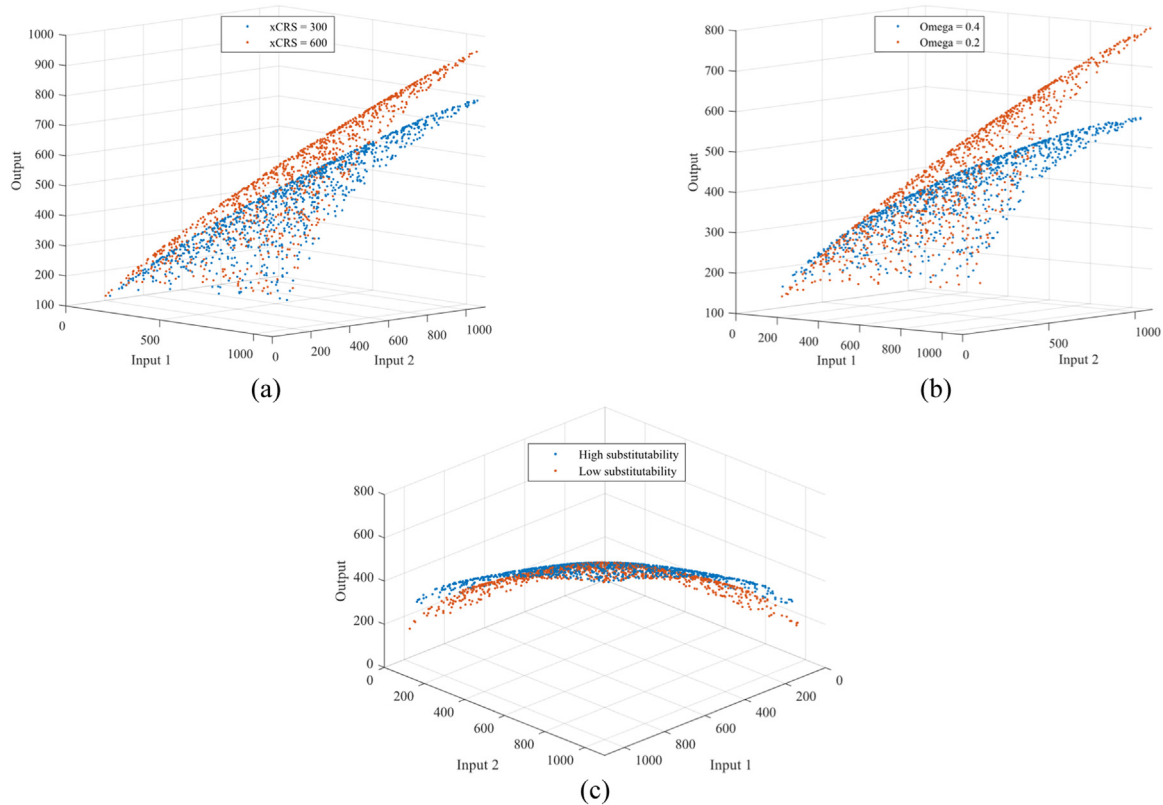


Fig. 2. Effect of x_i^{CRS} (a), ω (b), and ν (c) on the form of the production function.

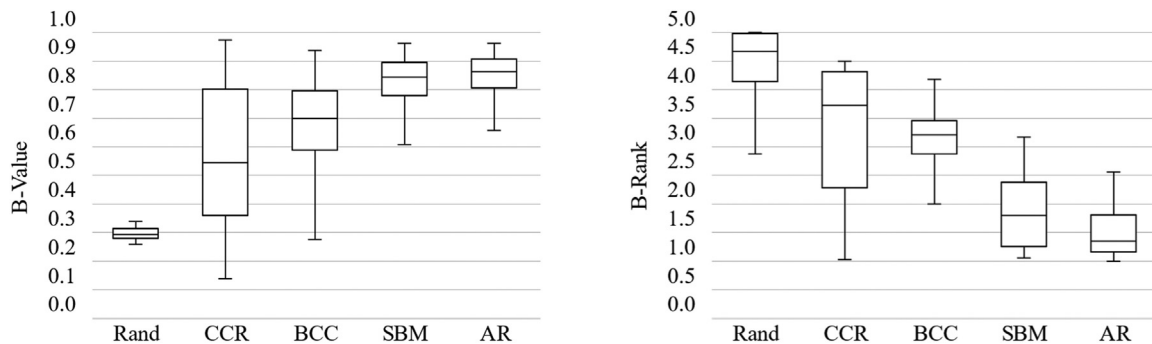


Fig. 3. Boxplots of B-Values and B-Ranks obtained from the models.

handle this issue, we use CORRI to represent the mean value of estimated inefficiencies within a margin of $\delta = 0.05$ around the true efficiencies. Using this indicator, the estimated efficiency of each model can be distinguished within 5% of its corresponding true efficiency. Compared to the basic DEA models, the AR and SBM models perform better. It is evident from the SPEAR indicator that the CCR model is barely able to mimic the true efficiency scores. In contrast, the AR and SBM indicate acceptable results. TOP and IN-EFF indicators provide the same result: AR and SBM exhibit high quality and outperform other models.

On the basis of Fig. 3, the accuracy of the VRS DEA models can be explained as follows. In the first place, the AR and SBM models perform significantly better than the BCC model, while it is the most popular model in DEA applications. BCC has a mean and StD of 0.649 and 0.201, respectively, indicating superior quality to CCR (mean of 0.574 and StD of 0.238) which is not surprising since our DGP is implemented using the VRS setting. However, it is a clear indication of the reliability of the results of the DGP and provides insight into the mechanism by which it operates. In terms

of the StD of the B-Values, the SBM and AR exhibit less dispersion from the corresponding mean values than the basic CCR and BCC DEA models. In light of the high B-Values for AR and SBM, which are close to 1.0, it can be said that these two models provide (nearly) accurate estimates. This result becomes even more significant when considering that these results represent the average over at least 50 replications of each scenario. Conversely, an examination of the minimum B-Values sheds some light on the vulnerable performances of all four models in some scenarios. Both SBM and AR models that have a minimum B-Value of 0.150 are performing better than the basic DEA models. The B-Rank, whose best value is equal to 1.0, is in agreement with the majority of certain findings testified by the B-Value. This indicator is not only a measure of dominance at the average level of scenarios but also takes into account every performance indicator in each replication. Overall, the AR model (with mean and StD of B-Rank of 1.479 and 0.354, respectively) performs marginally better than the SBM model (mean and StD of 1.877 and 0.568) and significantly better than the basic DEA models.

Table 4
Statistical values of performance indicators calculated for each model under the VRS setting.

Indicator	Statistics	Rand	CCR DEA	BCC DEA	AR DEA ($k = 2$)	SBM DEA
1-MAE	Max	0.866	0.987	0.984	0.986	0.986
	Min	0.782	0.245	0.376	0.253	0.257
	Mean	0.824	0.764	0.836	0.904	0.898
	StD	0.030	0.191	0.122	0.133	0.130
Rank (1-MAE)	Max	5.000	5.000	4.318	4.318	4.441
	Min	1.000	1.000	1.042	1.000	1.000
	Mean	3.901	3.625	3.257	1.809	2.379
	StD	1.290	1.621	0.705	0.712	0.830
SPEAR	Max	0.048	0.996	0.971	0.987	0.987
	Min	-0.041	0.057	-0.054	0.077	0.067
	Mean	0.000	0.634	0.703	0.858	0.841
	StD	0.011	0.269	0.277	0.186	0.188
Rank (SPEAR)	Max	5.000	4.154	4.711	2.422	3.077
	Min	3.244	1.000	2.339	1.000	1.018
	Mean	4.922	3.096	3.498	1.408	1.979
	StD	0.220	1.056	0.525	0.325	0.507
TOP	Max	0.198	0.924	0.885	0.905	0.905
	Min	0.118	0.155	0.133	0.158	0.155
	Mean	0.154	0.448	0.616	0.695	0.692
	StD	0.010	0.218	0.184	0.181	0.183
Rank (TOP)	Max	5.000	4.359	4.351	3.170	3.244
	Min	2.206	1.014	1.110	1.000	1.000
	Mean	4.681	3.314	2.612	1.447	1.540
	StD	0.538	0.963	0.562	0.406	0.471
INEFF	Max	0.193	0.972	0.910	0.973	0.973
	Min	0.117	0.168	0.123	0.188	0.187
	Mean	0.154	0.617	0.598	0.835	0.821
	StD	0.010	0.211	0.207	0.159	0.160
Rank (INEFF)	Max	5.000	4.083	4.531	1.868	2.656
	Min	2.580	1.048	1.706	1.000	1.000
	Mean	4.834	2.891	3.287	1.133	1.329
	StD	0.357	0.923	0.553	0.157	0.332
CORRI	Max	0.428	0.999	0.969	0.986	0.986
	Min	0.269	0.008	0.021	0.005	0.006
	Mean	0.344	0.405	0.494	0.737	0.707
	StD	0.053	0.333	0.279	0.265	0.264
Rank (CORRI)	Max	5.000	4.154	4.711	2.422	3.077
	Min	3.244	1.000	2.339	1.000	1.018
	Mean	4.922	3.096	3.498	1.408	1.979
	StD	0.220	1.056	0.525	0.325	0.507

Table 5
Results of conducting hypothesis tests.

Model	Number of Rejected Scenarios (%)
Rand	0 (0%)
CCR	3930 (50.5%)
BCC	2368 (30.5%)
AR	870 (11.2%)
SBM	829 (10.7%)

3.3. Results of hypothesis tests for comparing efficiency

The results of the statistical tests evaluating the null hypothesis that there is no difference in the distributions of true efficiency and estimated efficiency determined by the four VRS DEA models are presented in this section. The test statistic and critical value are calculated for each scenario, and if the test statistic is greater than the critical value, the null hypothesis is rejected. In Table 5, we report the distribution of the rejected scenarios. The value of 0.0 reported for the Rand can serve as a valid indicator of the robustness of the hypothesis tests conducted. This value is equal to 0 because both the true and estimated efficiencies by Rand are generated from the same distribution function. These findings also corroborate the main conclusions drawn from analyzing performance indicators. The total number of rejected scenarios in the AR and

SBM models (870 and 829, respectively) is considerably less than in the basic DEA models. Moreover, only 10% (11%) of scenarios have efficiency scores that are different from their true efficiency as calculated by the SBM (AR) model. By examining the rejected scenarios in more detail (see Tables E4 and E5 in Appendix E), it is apparent that the majority of them have fewer DMUs and more inputs. Moreover, these results underscore the importance of selecting the right RTS. This is because on average, the CCR DEA model fails to estimate the efficiency scores of 50% of scenarios generated under the VRS setting. BCC, which has been widely used in the DEA literature, is unquestionably outperformed by the AR and SBM models under the VRS setting.

In addition to $k = 2$, we set $k = 3, 4$ in the AR model to study the effect of the AR weight restrictions on the quality of efficiency estimates. The results of B-Value and B-Rank obtained from the AR model with $k = 3, 4$ against $k = 2$ are presented in Fig. 4. The medians are all at the same level. Therefore, it can be concluded that the AR models perform under setting different k 's almost identical. However, the box plots in these examples show relatively different distributions of B-Values and B-Ranks. The total number of rejected scenarios in the AR models with $k = 3, 4$ are 944 (12.1%) and 930 (11.9%), respectively, which are still considerably less than those obtained from the basic DEA models. The main reason for the better performance of the AR model is that the weights of inputs and outputs obtained from the basic DEA models are freely

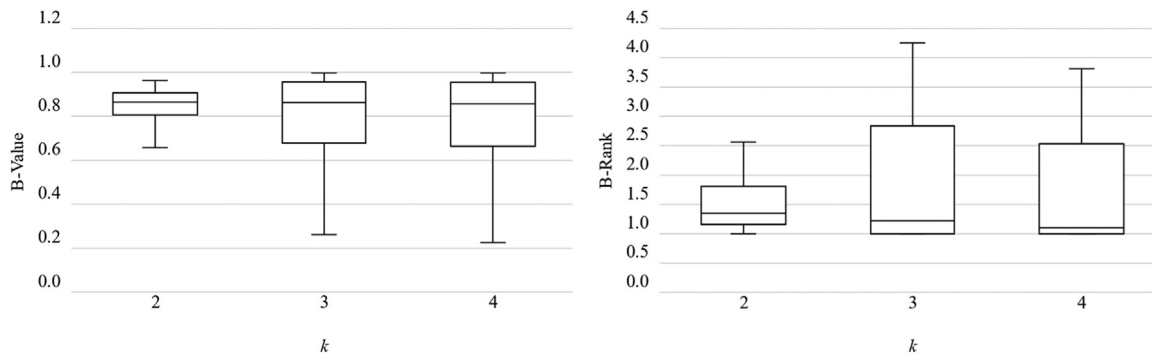


Fig. 4. Boxplots of B-Values and B-Ranks obtained from AR model where $k = 2, 3, 4$.

chosen. Therefore, they can get zero, which means they are excluded from the production possibilities set. This issue is evaded in AR models by restricting the weights.

3.4. Analysis of characteristics considered in the DGP

The purpose of this section is to investigate the identification of trends and patterns prompted by the ten different characteristics considered in the DGP. In Appendix E, we provide the descriptive statistics of the aggregated performance indicators and hypothesis tests according to the various values/levels defined for each characteristic. Based on the main drivers of these results, several consistency patterns emerge. Studies indicate that the size of the dataset, i.e., the number of DMUs and inputs, has a significant effect on the accuracy of DEA models. As reported subsequently, the results of our study confirm that increasing the size of the dataset results in decreasing the mean B-Values and in increasing the rejections. These two characteristics, however, are not the only ones responsible for the distinct influences. The use of more inputs and a low number of DMUs both negatively affect the mean B-Value. This results in more rejected scenarios as well. The mean B-Value of the BCC DEA model (see Table E3 in Appendix E) is reduced by 25% from 0.750 to 0.560 when we use 7 inputs instead of 2 and the number of rejections is almost doubled from 529 to 1134.

The lower bounds of 0.25 (low), 0.40 (medium), and 0.55 (high) for true efficiency levels reflect the fact that units with extremely poor efficiency cannot survive in the real world. B-Values and the number of rejected scenarios reported in Appendix E can be used to determine how true efficiency levels affect the quality of the DEA models. Increasing the lower bounds of true efficiencies causes a slight decline in the mean B-Values and a slight rise in the number of rejections in the DEA models. The quality of the DEA models is marginally diminishing by allocating a larger share of DMUs to the true efficiency frontier (efficiency score of 1.0). This may be partly explained by the fact that scaling down the lower bounds of the true efficiency results in a broader range of scores. Resulting in more DMUs are moving closer to the efficiency frontier. Due to this, the discrimination power of DEA models is reduced, while the negative effects are marginally present. When the importance of every input is different (ASYM), we see that the mean B-Values of all DEA models are to some extent less than when all inputs have equal importance in the production function. Accordingly, fewer scenarios are rejected under the SYM setting than under ASYM. According to our results, DEA estimations are not affected significantly by input importance.

Taking a look at the input substitution distribution, it is evident that when the input substitution is considered unequal, the performance of all DEA models is significantly better than when it is equal. In reality, substitution between all inputs utilized by DMUs does not need to be identical. The situation is different when in-

puts differ in substitutability. The AR and SBM DEA models are almost insensitive to substitutability variations. The high input substitutability adversely affects the performance of basic DEA models (CCR and BCC). Another two characteristics that are crucial to the form of the production function are the efficient size (x_i^{CRS}) and the extent of scale effects ω . In Appendix E, we demonstrate that when the efficient size is near the lower bound of the input range, i.e., 300, the performance of the VRS DEA models is marginally reduced since the scale effect starts at the beginning of the production function. As expected, this reduction in performance is more apparent in the CCR model. When the extent of the scale effect is increased, the performance of the basic DEA models CCR and BCC is diminished as the B-Values decrease and the number of rejected scenarios increases substantially. Once again, AR and SBM models perform better when the curvature of the production function is increased by increasing the extent of scale effects. Across all models, it is evident that larger input ranges result in less satisfactory results. This is very well reflected in the substantial increase in rejected scenarios. The results also reveal the trivial influence of the correlation of inputs upon the results of all DEA models. In real life, it is likely that there is a strong correlation between inputs, and that a complete lack of correlation is unlikely.

In summary, this set of results leads to a soundly clear ranking of the DEA models: $AR \approx SBM > BCC > CCR$. As a result of comparing the superior SBM and AR models, it is evident that despite almost identical B-Values and the number of rejections, some differences exist on the performance indicator level. Additionally, the results of B-Rank confirm the dominance of the AR model over the SBM model. The SBM model, however, shows almost the same performance as the AR model. The usage of both AR and SBM models as standard VRS DEA models can therefore be endorsed. However, the quality of the AR model might (strongly) depend on the setting of the weight restrictions and special attention should be given.

4. Conclusions

In this paper, we propose a method based on Monte Carlo simulation to assess the quality of DEA model estimates. Our method involves generating data by using a flexible technology (Translog production function) that satisfies microeconomic regularity conditions such as convexity and monotonicity. Prior studies have lacked diversity in the DGPs, which is a serious handicap when evaluating the quality of DEA model estimations. We generate 7776 distinct scenarios under the VRS setting by defining a variety of characteristics. Our evaluations of the quality of estimates obtained from DEA models are based on five performance indicators, as well as DEA-based hypothesis tests. Furthermore, we demonstrate how a valid range of characteristics and parameters can be derived when the necessary and sufficient microeconomic conditions are all met.

To our knowledge, this is the first study that compares the quality of VRS DEA models to date. We show that the BCC model, which is the most commonly used VRS DEA model in the literature, is outperformed by AR and SBM models. According to the hypothesis test's results, we find that more than 30% of BCC model estimations differ from the distribution of the true efficiency, but this rejection percentage is 11% for AR and 10% for SBM models. It is noteworthy that the AR model emerged at the top without applying any special tuning to the virtual weight restrictions. However, it may be too complex to explicitly articulate weights in some applications. We, therefore, endorse the establishment of the SBM model as the standard VRS DEA model in which there are no prior conditions to be comprehended on weights since its performance is almost equal to that of the AR model. From our perspective, the dominance of the AR and SBM models can be explained by the presence of slacks. While the BCC model ignores slacks entirely in reporting the efficiency score, the SBM model calculates the efficiency score directly based on the slacks. Furthermore, the AR model prevents the emergence of slacks by assigning boundaries to the weights. We also examine the impact of characteristics used for generating scenarios on the quality of the DEA estimates. According to our results, the most important factors affecting the quality of VRS DEA models are the number of inputs, range of inputs, distribution of input substitution, and scale effects. Our results may also be useful for decision-makers who might use them as a guideline for their own DEA studies to ensure acceptable results and accuracy.

Consideration of the single-output case is one of the limitations of our DGP. The methodology may therefore be generalized to meaningful multi-input multi-output cases in the future. The goal of our DGP is to generate artificial datasets that fulfill the monotonicity and curvature conditions for every single generated scenario. In order to secure a similar guarantee in the multi-output case, the formulation of the Translog production function needs to be updated as the importance of outputs and their substitutions must be considered. In particular, this means the necessary conditions for guaranteeing the VRS setting with respect to outputs such as imposing homogeneity of degree +1 and convexity in outputs need to be modeled and considered.

Furthermore, the proposed DGP identifies the deviation of the output from the efficiency frontier as a single inefficiency term. A stochastic framework is another method of extending the DGP. The DGP can then be extended by defining the inefficiency score as the sum of two terms: inefficiency and noise. Another line of investigation would be extending our method for panel data with a time trend. Using this, we can assess and improve the accuracy of Malmquist productivity index calculations and their decomposition. In addition, we believe the methodology presented here can also be used to investigate other multi-input multi-output production functions, such as the one presented by Färe, Grosskopf, Noh

and Weber (2005). All of this may eventually make DEA models more practical by increasing their reliability and showing how accurate their estimations are to decision-makers.

Appendix A. Performance indicators

In Table A, θ_j and $\hat{\theta}_j$ denote the true efficiency and efficiency score calculated by the DEA model for j th DMU ($j \in \{1, \dots, n\}$), respectively. There are two important points to consider when defining performance indicators. First, DEA estimates $\hat{\theta}_j = 1$ for some DMUs while their corresponding true efficiency scores obtained from the DGP might be less than (but close to) one, i.e., $\theta_j = 0.90 < 1.0$ since they are based on a random continuous function. Second, in the small-size samples, it is expected that only a few DMUs (or no DMU) with a true efficiency score of 1.0 have been produced. These two points preclude using a simple indicator that only evaluates whether DMUs with an estimated efficiency score of 1.0 ($\hat{\theta}_j = 1.0$) also have a corresponding true efficiency score of 1.0 ($\theta_j = 1.0$). Our objective is therefore to determine whether the DEA models are capable of identifying the top-performing DMUs in a sample, although, not all of them have a true efficiency score of 1.0 but are close to it. In light of these two points, TOP and IN-EFF are performance indicators based on the quantiles of worst- and best-performing DMUs, respectively. This study defines an efficient DMU as one that has at least as high a true efficiency value as a specific quantile ($Q(\varepsilon)$) of the distribution of true efficiency. In the same manner, a DMU is inefficient if and only if its true efficiency is less than or equal to $Q(1 - \varepsilon)$. For example, consider 50 DMUs ($n = 50$) where $\varepsilon = 0.8$. In the ascending order of true efficiencies, $Q(\varepsilon) = \theta_j$ where $j = 40$. The same logic can be applied to $Q(1 - \varepsilon)$. In this way, we can handle multiple efficiency distributions in the DGP as well as compare different scenarios. Ideally, parameter ε should be large enough to serve as a satisfactory limit for efficient DMUs. We also employ the CORRI to track the mean value of estimates in certain corridors around the true efficiencies, since MAE cannot provide information on the deviation.

The parameters δ and γ determine the tightness of the corridors and the number of corridors, respectively. As in Kohl and Brunner (2020), we also use a corrugated line of $\delta = 0.05$ to test an estimated model's efficacy at most 5% points. This is in addition to the corresponding true score. Having generated the data (including inputs, outputs, and a true efficiency score) of a scenario and calculated the efficiency scores by DEA models, we constructed the performance indicators. To aggregate and represent all the performance indicators with a single score we use B-Value. To capture the influence of dominance, we also introduce a second aggregated indicator called B-Rank. In Table A, the last two rows give the formulas for these two aggregated indicators.

Table A
Performance indicators used for quality evaluation of DEA models (Kohl & Brunner, 2020).

Indicator	Symbol	Formula
Mean absolute error	MAE	$\frac{1}{n} \sum_{j=1}^n \theta_j - \hat{\theta}_j $
Spearman Correlation Coefficient	SPEAR	$\frac{\sum_j (\text{Rg}(\hat{\theta}_j) - \overline{\text{Rg}(\hat{\theta})})(\text{Rg}(\theta_j) - \overline{\text{Rg}(\theta)})}{\sqrt{\sum_j (\text{Rg}(\hat{\theta}_j) - \overline{\text{Rg}(\hat{\theta})})^2} \sqrt{\sum_j (\text{Rg}(\theta_j) - \overline{\text{Rg}(\theta)})^2}}$
Best-performing DMUs	TOP	$\frac{ \{j: \theta_j \geq Q(\varepsilon)\} }{ \{j: \hat{\theta}_j \geq Q(\varepsilon)\} } \cdot (1 - \frac{\max\{ \{j: \hat{\theta}_j \geq Q(\varepsilon)\} - \{j: \theta_j \geq Q(\varepsilon)\} \}}{n})$
Worst-performing DMUs	INEFF	$\frac{ \{j: \theta_j \leq Q(1-\varepsilon)\} }{ \{j: \hat{\theta}_j \leq Q(1-\varepsilon)\} } \cdot (1 - \frac{\max\{ \{j: \hat{\theta}_j \leq Q(1-\varepsilon)\} - \{j: \theta_j \leq Q(1-\varepsilon)\} \}}{n})$
Mean value over the results of the corridor	CORRI	$\sum_{k=1}^{\gamma} \frac{1}{\gamma} \frac{ \{j: \theta_j - \hat{\theta}_j \leq k\delta\} }{n}$
Benchmark value	B-Value	$\frac{(1-\text{MAE})+\text{SPEAR}+\text{EFF}+\text{INEFF}+\text{CORRI}}{5}$
Benchmark rank	B-Rank	$\frac{\text{rank}(1-\text{MAE}) + \text{rank}(\text{SPEAR}) + \text{rank}(\text{EFF}) + \text{rank}(\text{INEFF}) + \text{rank}(\text{CORRI})}{5}$

Appendix B. Propositions and proofs

Proposition 1. The definition provided in Eq. (3) for α_i fulfills the condition of $\sum_{i \in \mathcal{M}} \alpha_i > 1$.

Proof. We need to prove that the definition provided for α_i in Eq. (3) guarantees the implementation of the first condition of the VRS regime. Mathematically speaking, $\sum_{i=1}^m \alpha_i = \sum_{i=1}^m \frac{1+\omega}{m} = \sum_{i=1}^m (\frac{1}{m} + \frac{\omega}{m}) = (m \cdot \frac{1}{m} + m \cdot \frac{\omega}{m}) = 1 + \omega \rightarrow \omega > 0 \rightarrow \sum_{i=1}^m \alpha_i > 1$. ■

Proposition 2. The definition provided in Eq. (4) for α_i fulfills the condition of $\sum_{i \in \mathcal{M}} \alpha_i > 1$.

Proof. We show that the definition provided for α_i in Eq. (4) respects the first condition of the VRS, i.e., $\sum_{i=1}^m \alpha_i = \sum_{i=1}^m \frac{(1+\omega) \cdot (i+m)}{1.5m^2+0.5m} = \sum_{i=1}^m \frac{(1+\omega) \cdot i}{1.5m^2+0.5m} + \sum_{i=1}^m \frac{(1+\omega) \cdot m}{1.5m^2+0.5m} = (1+\omega) \cdot [\frac{\frac{1}{2}m \cdot (m+1)}{1.5m^2+0.5m} + \frac{m \cdot m}{1.5m^2+0.5m}] = (1+\omega) \cdot [\frac{1.5m^2+0.5m}{1.5m^2+0.5m}] = 1 + \omega \rightarrow \omega > 0 \rightarrow \sum_{i=1}^m \alpha_i > 1$. ■

Proposition 3. Definitions provided in Eq. (6) fulfill $\beta_{ii} + \sum_{h \in \mathcal{M} \setminus \{i\}} \beta_{ih} = -\frac{\omega}{m \cdot \ln x_i^{CRS}}$, $\forall i$.

Proof. By replacing β_{ii} and β_{ih} in $\beta_{ii} + \sum_{h \neq i} \beta_{ih}$ and operating it, we have

$$\beta_{ii} + \sum_{h \in \mathcal{M} \setminus \{i\}} \beta_{ih} = \frac{-v \cdot \omega}{m \cdot \ln x_i^{CRS}} + \sum_{h \neq i} \frac{(v-1) \cdot \omega}{m \cdot (m-1) \ln x_i^{CRS}} = \frac{-v \cdot \omega}{m \cdot \ln x_i^{CRS}} + \frac{(m-1)(v-1) \cdot \omega}{m \cdot (m-1) \ln x_i^{CRS}} = -\frac{\omega}{m \cdot \ln x_i^{CRS}} \rightarrow \beta_{ii} + \sum_{h \neq i} \beta_{ih} = -\frac{\omega}{m \cdot \ln x_i^{CRS}}, \forall i. \blacksquare$$

Proposition 4. Definitions provided in Eq. (8) fulfill Eq. (5).

Proof. We call the unequal substitution distribution defined by Kohl and Brunner (2020), i.e., $\sigma'_{ii} = -\frac{m \cdot (1.5 - \frac{i-1}{m-1}) \cdot (2-2 \cdot \frac{i-1}{m-1})}{1.5 \cdot m - 2}$, $\forall i$ and $\sigma'_{ih} = \frac{2 - \frac{h-1}{m-1} - \frac{i-1}{m-1}}{1.5 \cdot m - 2}$, $\forall \{i, h | i \neq h\}$. From the proof provided by them, we know that $\sigma'_{ii} + \sum_{h \in \mathcal{M} \setminus \{i\}} \sigma'_{ih} = 0, \forall i$.⁷ Now, by replacing these two expressions in Eq. (5) and operating, we have: $\sum_{i \in \mathcal{M}} (\beta_{ii} + \sum_{h \in \mathcal{M} \setminus \{i\}} \beta_{ih}) =$

$$\begin{aligned} & \sum_{i \in \mathcal{M}} \left(-\frac{\omega \cdot (1-v \cdot \sigma'_{ii})}{m \cdot \ln x_i^{CRS}} + \sum_{h \in \mathcal{M} \setminus \{i\}} \frac{\omega \cdot v}{m \cdot \ln x_i^{CRS}} \cdot \sigma'_{ih} \right) = \\ & \sum_{i \in \mathcal{M}} \left(-\frac{\omega}{m \cdot \ln x_i^{CRS}} + \frac{\omega \cdot v \cdot \sigma'_{ii}}{m \cdot \ln x_i^{CRS}} + \frac{\omega \cdot v}{m \cdot \ln x_i^{CRS}} \cdot \sum_{h \in \mathcal{M} \setminus \{i\}} \sigma'_{ih} \right) \\ & = \sum_{i \in \mathcal{M}} \left(-\frac{\omega}{m \cdot \ln x_i^{CRS}} \cdot (1 + v \cdot \sigma'_{ii} + v \cdot \sum_{h \in \mathcal{M} \setminus \{i\}} \sigma'_{ih}) \right) = \\ & \sum_{i \in \mathcal{M}} \left(-\frac{\omega}{m \cdot \ln x_i^{CRS}} \cdot (1 + v \cdot (\sigma'_{ii} + \sum_{h \in \mathcal{M} \setminus \{i\}} \sigma'_{ih})) \right) = \\ & \sum_{i \in \mathcal{M}} \left(-\frac{\omega}{m \cdot \ln x_i^{CRS}} \cdot (1 + 0) \right) \\ & = \sum_{i \in \mathcal{M}} \left(-\frac{\omega}{m \cdot \ln x_i^{CRS}} \cdot (1 + 0) \right) = \sum_{i \in \mathcal{M}} -\frac{\omega}{m \cdot \ln x_i^{CRS}} = -\frac{\omega}{\ln x^{CRS}}. \blacksquare \end{aligned}$$

Appendix C. DEA models under evaluation

The basic DEA model (known as CCR) was introduced by Charnes et al. (1978). They define a measure of efficiency by maximizing the ratio of the weighted sum of outputs over the weighted sum of inputs for each DMU. Consider input vector $\mathbf{X} = (x_{10}, \dots, x_{m0})$ and output vector $\mathbf{Y} = (y_{10}, \dots, y_{s0})$, then the relative efficiency of $DMU_j \forall j = 1, \dots, n$ can be formulated as $TE_j = \frac{\sum_{r=1}^s u_r y_{rj}}{\sum_{i=1}^m v_i x_{ij}}$ where, u_r and v_i are the weights of output r and input i , respectively. The mathematical formulation of the CCR DEA model can be presented as follows:

$$\max \theta_o \tag{C.1}$$

$$\text{s.t. } \sum_{j=1}^n x_{ij} \lambda_j \leq x_{i0}, \forall i \tag{C.2}$$

$$\sum_{j=1}^n y_{rj} \lambda_j \geq \theta_o y_{r0}, \forall r \tag{C.3}$$

$$\lambda_j \geq 0, \forall j \tag{C.4}$$

where, θ_o shows the technical efficiency of DMU_o and λ_j are the intensity variables. In the CCR DEA model, the assumption of CRS is underlined. The BCC (Banker-Charnes-Cooper) model (Banker et al., 1984) is the most representative extension of the CCR DEA model in which VRS technology is accommodated. The BCC DEA model can be formed by adding the convexity constraint $\sum_{j=1}^n \lambda_j = 1$ to the CCR DEA model (C).

There might be many zeros in the optimal weights of the CCR and BCC models, indicating that the evaluating DMU may have a weakness in the factors (inputs and outputs) compared to the efficient DMUs. Having no control over the boundaries of optimal weights leads to the emerging AR DEA model, which constrains the weight of special inputs/outputs relative to others (Thompson, Singleton, Thrall & Smith, 1986). In the literature (see Allen, Athanassopoulos, Dyson and Thanassoulis (1997) and Pedraja-Chaparro et al. (1997)), several approaches have been developed to restrict the weights of DEA models. There are two dimensions for the weight restrictions in the AR models. The first dimension relates to the weights to be constrained i.e., raw or virtual weight restriction. The second dimension is about the limits placed on the weights which can have either absolute or relative weight restrictions. A raw weight restriction limits just the weight within the primal multiplier model, whereas a virtual weight restriction limits the product of weight and input, i.e., $v_i \cdot x_{ij}$. The ratio between two weights is affected by relative restrictions, as opposed to absolute restrictions, which affect only one weight. The focus of our analysis is on relative weight restrictions following Pedraja-Chaparro et al. (1997). Since we deal with different input elasticities, we chose to apply virtual weight restrictions. To determine whether the quality of the AR model is affected by weight restrictions, we consider relative virtual weight constraints by setting $k = 2, 3, 4$ in our computational study (see Section 3.3).

$$\min \sum_{i=1}^m v_i x_{i0} \tag{C.5}$$

$$\text{s.t. } \sum_{i=1}^m v_i x_{ij} - \sum_{r=1}^s u_r y_{rj} \geq 0, \forall j \tag{C.6}$$

$$\sum_{r=1}^s u_r y_{r0} = 1 \tag{C.7}$$

$$u_r, v_i \geq 0, \forall r, i \tag{C.8}$$

$$\frac{v_i x_{i0}}{v_h x_{h0}} \leq k, \forall i, h, i \neq h \tag{C.9}$$

Both CCR and BCC DEA models are radial where inputs are proportionally reduced, and outputs are proportionally expanded. This assumption can be restrictive. For example, when labor, capital, and material are employed as inputs, some of them may not change proportionally and may be substituted. A further shortcoming of radial models is that they do not consider slacks when reporting efficiency scores. There are often loads of non-radial slacks left. These limitations lead to the expansion of non-radial models. SBM DEA is a non-radial model that deals directly with slacks in reporting efficiency scores (Tone, 2001). The non-oriented SBM DEA model under the VRS setting is a non-linear model that can be reformulated as a linear counterpart by using Charnes–Cooper transformation approach (Charnes & Cooper, 1962) as follows:

$$\min \rho_o = t - \frac{1}{m} \sum_{i=1}^m \frac{s_i^-}{x_{i0}} \tag{C.10}$$

⁷ A detailed derivation of σ'_{ii} and σ'_{ih} can be found in Kohl and Brunner (2020).

Table D
One scenario (50 DMUs, two inputs, and one output) generated by the developed DGP.

DMU	Input 1	Input 2	Output 1	True Eff.	DMU	Input 1	Input 2	Output 1	True Eff.
1	996.00	722.00	1147.38	0.7879	26	234.00	610.00	474.28	0.7814
2	295.00	964.00	641.16	0.7908	27	259.00	1015.00	629.63	0.8469
3	122.00	997.00	215.45	0.5526	28	985.00	974.00	1254.82	0.7422
4	863.00	173.00	264.04	0.5058	29	894.00	583.00	1172.46	0.9484
5	307.00	948.00	821.23	0.9880	30	816.00	978.00	621.86	0.4032
6	1092.00	122.00	372.20	0.9502	31	989.00	814.00	1197.65	0.7737
7	143.00	565.00	323.09	0.7819	32	961.00	362.00	528.62	0.5639
8	1045.00	310.00	495.89	0.5783	33	628.00	1002.00	695.69	0.5156
9	1075.00	573.00	933.06	0.7117	34	249.00	132.00	233.94	0.7640
10	102.00	832.00	219.42	0.6729	35	1051.00	939.00	1708.48	0.9983
11	514.00	399.00	544.08	0.6858	36	808.00	962.00	970.23	0.6369
12	812.00	151.00	424.62	0.9202	37	749.00	638.00	1107.77	0.9195
13	814.00	724.00	792.67	0.5937	38	390.00	862.00	685.65	0.7191
14	535.00	228.00	383.96	0.6706	39	939.00	867.00	1383.95	0.8863
15	227.00	311.00	295.18	0.6347	40	997.00	149.00	268.93	0.5748
16	146.00	1058.00	365.69	0.7916	41	923.00	995.00	1384.13	0.8365
17	906.00	534.00	1095.83	0.9283	42	258.00	380.00	520.20	0.9540
18	813.00	715.00	1023.81	0.7721	43	765.00	835.00	1250.16	0.9000
19	783.00	686.00	738.67	0.5788	44	709.00	359.00	606.92	0.7170
20	230.00	418.00	515.33	0.9751	45	985.00	954.00	1316.94	0.7868
21	1091.00	826.00	1219.96	0.7496	46	773.00	1079.00	1494.10	0.9561
22	243.00	275.00	346.09	0.7584	47	356.00	991.00	468.93	0.5017
23	1093.00	674.00	1141.91	0.7857	48	660.00	297.00	315.07	0.4310
24	651.00	992.00	1112.20	0.8106	49	111.00	613.00	198.47	0.5828
25	970.00	1025.00	1364.69	0.7938	50	1090.00	1083.00	1573.86	0.8430

$$s.t.t + \frac{1}{s} \sum_{r=1}^s \frac{s_r^+}{y_{ro}} = 1 \tag{C.11}$$

$$t \cdot x_{io} = \sum_{j=1}^n x_{ij} \lambda_j + s_i^-, \quad \forall i \tag{C.12}$$

$$t \cdot y_{ro} = \sum_{j=1}^n y_{rj} \lambda_j - s_r^+, \quad \forall r \tag{C.13}$$

$$\sum_{j=1}^n \lambda_j = 1 \tag{C.14}$$

$$s_i^-, s_r^+, \lambda_j \geq 0, \quad \forall i, r, j \text{ and } t > 0 \tag{C.15}$$

where ρ_o is the SBM-efficiency, s^- and s^+ are the vector of input and output slacks, respectively. t is a positive scalar variable used during the transformation process. Consider the optimal solution system as of the non-oriented SBM DEA model be $\{\rho^*, t^*, \lambda^*, s^-, s^+\}$ then, the optimal solution of the SBM DEA model can be defined as $\{\rho^*, t^*, \lambda^*/t^*, s^-/t^*, s^+/t^*\}$.

Appendix D. One sample scenario

Table D.

Appendix E. Detailed results of analysis of characteristics

Table E1, Table E2

Table E1
Rand Model.

Model	Characteristic	Value/Level	B-Value				Rejection		
			Max	Min	Mean	StD	Mean	StD	Sum
Rand	True efficiency level	Low	0.299	0.259	0.276	0.005	0	0	0
		Medium	0.320	0.277	0.293	0.005	0	0	0
		High	0.339	0.301	0.317	0.005	0	0	0
#DMU	50	50	0.339	0.264	0.298	0.017	0	0	0
		150	0.332	0.259	0.294	0.017	0	0	0
		450	0.330	0.261	0.295	0.017	0	0	0
#Inputs	2	2	0.339	0.259	0.295	0.017	0	0	0
		5	0.335	0.262	0.295	0.017	0	0	0
		7	0.332	0.262	0.295	0.017	0	0	0
Input Importance	ASYM	ASYM	0.335	0.261	0.295	0.017	0	0	0
		SYM	0.339	0.259	0.295	0.017	0	0	0
Input substitution distribution	Unequal	Equal	0.335	0.262	0.295	0.017	0	0	0
		Unequal	0.339	0.259	0.295	0.017	0	0	0
Input substitutability	High	High	0.339	0.259	0.295	0.017	0	0	0
		Low	0.335	0.262	0.295	0.017	0	0	0
Efficient size	300	300	0.339	0.259	0.295	0.017	0	0	0
		600	0.335	0.262	0.295	0.017	0	0	0
		[100; 1100]	0.339	0.259	0.295	0.017	0	0	0
Input range	[100; 10,100]	[100; 10,100]	0.335	0.261	0.295	0.017	0	0	0
		0.2	0.339	0.262	0.295	0.017	0	0	0
Extent of scale effects	0.4	0.4	0.332	0.259	0.295	0.017	0	0	0
		0.8	0.332	0.262	0.296	0.017	0	0	0
		0	0.335	0.259	0.295	0.017	0	0	0
Input correlation	0.4	0	0.335	0.262	0.295	0.017	0	0	0
		0.4	0.335	0.262	0.295	0.017	0	0	0
		0.8	0.339	0.261	0.295	0.017	0	0	0

Table E2
CCR DEA Model.

Model	Characteristic	Value/Level	B-Value				Rejection		
			Max	Min	Mean	StD	Mean	StD	Sum
CCR	True efficiency level	Low	0.974	0.192	0.614	0.223	0.484	0.500	1254
		Medium	0.967	0.169	0.576	0.236	0.505	0.500	1310
		High	0.958	0.140	0.532	0.246	0.527	0.499	1366
#DMU	50	50	0.973	0.154	0.573	0.227	0.577	0.494	1495
		150	0.974	0.142	0.573	0.240	0.522	0.500	1352
		450	0.973	0.140	0.575	0.246	0.418	0.493	1083
#Inputs	2	2	0.965	0.164	0.627	0.232	0.402	0.490	1042
		5	0.934	0.140	0.571	0.238	0.406	0.491	1052
		7	0.974	0.236	0.523	0.233	0.708	0.455	1836
Input Importance	ASYM	ASYM	0.974	0.140	0.573	0.238	0.508	0.500	1977
		SYM	0.973	0.148	0.574	0.238	0.502	0.500	1953
Input substitution distribution	Equal	Equal	0.974	0.140	0.456	0.200	0.717	0.450	2789
		Unequal	0.973	0.240	0.691	0.214	0.293	0.455	1141
Input substitutability	High	High	0.974	0.146	0.629	0.234	0.409	0.492	1590
		Low	0.965	0.140	0.518	0.229	0.602	0.490	2340
Efficient size	300	300	0.973	0.140	0.553	0.236	0.540	0.498	2101
		600	0.974	0.153	0.594	0.238	0.470	0.499	1829
Input range	[100; 1100]	[100; 1100]	0.974	0.215	0.677	0.231	0.344	0.475	1337
		[100; 10,100]	0.892	0.140	0.471	0.197	0.667	0.471	2593
Extent of scale effects	0.2	0.2	0.974	0.236	0.667	0.218	0.341	0.474	883
		0.4	0.960	0.193	0.574	0.229	0.525	0.499	1360
		0.8	0.950	0.140	0.480	0.229	0.651	0.477	1687
Input correlation	0	0	0.973	0.148	0.576	0.232	0.461	0.499	1196
		0.4	0.974	0.142	0.574	0.238	0.513	0.500	1329
		0.8	0.973	0.140	0.572	0.243	0.542	0.498	1405

Table E3
BCC DEA Model.

Model	Characteristic	Value/Level	B-Value				Rejection		
			Max	Min	Mean	StD	Mean	StD	Sum
BCC	True efficiency level	Low	0.938	0.138	0.654	0.209	0.284	0.451	736
		Medium	0.928	0.154	0.650	0.201	0.307	0.462	797
		High	0.926	0.162	0.644	0.192	0.322	0.467	835
#DMU	50	50	0.889	0.148	0.634	0.184	0.391	0.488	1013
		150	0.923	0.141	0.652	0.204	0.365	0.482	946
		450	0.938	0.138	0.662	0.212	0.158	0.365	409
#Inputs	2	2	0.938	0.558	0.715	0.137	0.204	0.403	529
		5	0.880	0.138	0.673	0.202	0.272	0.445	705
		7	0.830	0.160	0.560	0.220	0.438	0.496	1134
Input Importance	ASYM	ASYM	0.938	0.141	0.647	0.203	0.310	0.463	1206
		SYM	0.935	0.138	0.652	0.199	0.299	0.458	1162
Input substitution distribution	Equal	Equal	0.938	0.138	0.556	0.232	0.462	0.499	1798
		Unequal	0.935	0.561	0.742	0.096	0.147	0.354	570
Input substitutability	High	High	0.938	0.138	0.688	0.213	0.196	0.397	762
		Low	0.876	0.141	0.611	0.179	0.413	0.492	1606
Efficient size	300	300	0.938	0.138	0.634	0.209	0.328	0.469	1274
		600	0.935	0.146	0.665	0.191	0.281	0.450	1094
Input range	[100; 1100]	[100; 1100]	0.938	0.196	0.688	0.174	0.210	0.408	818
		[100; 10,100]	0.931	0.138	0.611	0.217	0.399	0.490	1550
Extent of scale effects	0.2	0.2	0.938	0.423	0.739	0.119	0.162	0.369	420
		0.4	0.925	0.160	0.653	0.195	0.303	0.460	786
		0.8	0.875	0.138	0.556	0.228	0.448	0.497	1162
Input correlation	0	0	0.935	0.150	0.641	0.190	0.271	0.444	702
		0.4	0.938	0.143	0.651	0.202	0.311	0.463	807
		0.8	0.933	0.138	0.657	0.210	0.331	0.471	859

Table E4
AR DEA Model ($k = 2$).

Model	Characteristic	Value/Level	B-Value				Rejection		
			Max	Min	Mean	StD	Mean	StD	Sum
AR	True efficiency level	Low	0.962	0.208	0.812	0.165	0.109	0.311	282
		Medium	0.963	0.185	0.806	0.179	0.111	0.315	289
		High	0.962	0.150	0.798	0.193	0.115	0.320	299
	#DMU	50	0.907	0.164	0.775	0.165	0.124	0.330	322
		150	0.947	0.152	0.812	0.181	0.114	0.318	296
		450	0.963	0.150	0.831	0.188	0.097	0.296	252
	#Inputs	2	0.963	0.854	0.919	0.029	0.000	0.000	0
		5	0.931	0.216	0.796	0.160	0.103	0.305	268
		7	0.923	0.150	0.703	0.216	0.232	0.422	602
	Input Importance	ASYM	0.963	0.150	0.799	0.184	0.121	0.327	472
		SYM	0.962	0.160	0.812	0.175	0.102	0.303	398
	Input substitution distribution	Equal	0.962	0.150	0.746	0.233	0.224	0.417	870
		Unequal	0.963	0.687	0.866	0.056	0.000	0.000	0
	Input substitutability	High	0.963	0.158	0.806	0.180	0.112	0.315	435
		Low	0.961	0.150	0.806	0.179	0.112	0.315	435
	Efficient size	300	0.962	0.150	0.795	0.192	0.124	0.330	482
		600	0.963	0.175	0.816	0.165	0.100	0.300	388
		[100; 1100]	0.963	0.284	0.846	0.123	0.055	0.229	215
	Input range	[100; 10,100]	0.954	0.150	0.765	0.214	0.168	0.374	655
		0.2	0.962	0.681	0.869	0.056	0.000	0.000	0
	Extent of scale effects	0.4	0.963	0.280	0.823	0.143	0.086	0.280	223
0.8		0.958	0.150	0.726	0.250	0.250	0.433	647	
0		0.961	0.160	0.789	0.176	0.102	0.303	265	
Input correlation	0.4	0.963	0.152	0.808	0.179	0.117	0.322	304	
	0.8	0.962	0.150	0.820	0.182	0.116	0.320	301	

Table E5
SBM DEA Model.

Model	Characteristic	Value/Level	B-Value				Rejection		
			Max	Min	Mean	StD	Mean	StD	Sum
SBM	True efficiency level	Low	0.962	0.206	0.798	0.165	0.101	0.301	262
		Medium	0.963	0.185	0.793	0.178	0.107	0.309	278
		High	0.962	0.150	0.786	0.192	0.111	0.315	289
	#DMU	50	0.907	0.165	0.759	0.165	0.121	0.326	314
		150	0.947	0.153	0.799	0.180	0.107	0.309	277
		450	0.963	0.150	0.818	0.186	0.092	0.289	238
	#Inputs	2	0.963	0.854	0.919	0.029	0.000	0.000	0
		5	0.908	0.213	0.781	0.154	0.096	0.294	248
		7	0.881	0.150	0.676	0.204	0.224	0.417	581
	Input Importance	ASYM	0.963	0.150	0.788	0.183	0.114	0.318	443
		SYM	0.962	0.161	0.796	0.175	0.099	0.299	386
	Input substitution distribution	Equal	0.962	0.150	0.737	0.231	0.213	0.410	829
		Unequal	0.963	0.645	0.847	0.068	0.000	0.000	0
	Input substitutability	High	0.963	0.159	0.793	0.179	0.106	0.308	413
		Low	0.961	0.150	0.791	0.178	0.107	0.309	416
	Efficient size	300	0.962	0.150	0.782	0.191	0.117	0.322	456
		600	0.963	0.175	0.802	0.165	0.096	0.295	373
		[100; 1100]	0.963	0.283	0.828	0.126	0.049	0.216	191
	Input range	[100; 10,100]	0.954	0.150	0.756	0.213	0.164	0.370	638
		0.2	0.962	0.645	0.851	0.066	0.000	0.000	0
	Extent of scale effects	0.4	0.963	0.279	0.809	0.144	0.079	0.271	206
0.8		0.958	0.150	0.716	0.248	0.240	0.427	623	
0		0.962	0.161	0.776	0.177	0.094	0.292	244	
Input correlation	0.4	0.963	0.153	0.794	0.179	0.112	0.315	290	
	0.8	0.962	0.150	0.806	0.180	0.114	0.318	295	

References

- Allen, R., Athanassopoulos, A., Dyson, R. G., & Thanassoulis, E. (1997). Weights restrictions and value judgements in data envelopment analysis: Evolution, development and future directions. *Annals of Operations Research*, 73, 13–34.
- Balk, B. M. (2001). Scale efficiency and productivity change. *Journal of Productivity Analysis*, 15(3), 159–183.
- Banker, R., Natarajan, R., & Zhang, D. (2019). Two-stage estimation of the impact of contextual variables in stochastic frontier production function models using Data Envelopment Analysis: Second stage OLS versus bootstrap approaches. *European Journal of Operational Research*, 278(2), 368–384.
- Banker, R. D. (1993). Maximum likelihood, consistency and data envelopment analysis: A statistical foundation. *Management Science*, 39(10), 1265–1273.
- Banker, R. D., Charnes, A., & Cooper, W. W. (1984). Some models for estimating technical and scale inefficiencies in data envelopment analysis. *Management Science*, 30(9), 1078–1092.
- Banker, R. D., & Natarajan, R. (2011). Statistical tests based on DEA efficiency scores. In W. W. Cooper, L. M. Seiford, & J. Zhu (Eds.), *Handbook on data envelopment analysis* (eds) (pp. 273–295). Boston, MA: Springer US.
- Banker, R. D., Zheng, Z., & Natarajan, R. (2010). DEA-based hypothesis tests for comparing two groups of decision making units. *European Journal of Operational Research*, 206(1), 231–238.
- Bogetoft, P., & Otto, L. (2011a). Additional Topics. In Bogetoft P. SFA, & L. Otto (Eds.), *Benchmarking with DEA, SFA, and R* (eds) (pp. 233–262). New York, NY: Springer New York.
- Bogetoft, P., & Otto, L. (2011b). Statistical Analysis. In Bogetoft P. DEA, & L. Otto (Eds.), *Benchmarking with DEA, SFA, and R* (eds) (pp. 155–196). New York, NY: Springer New York.
- Charnes, A., & Cooper, W. W. (1962). Programming with linear fractional functionals. *Naval Research Logistics*, 9(3–4), 181–186.
- Charnes, A., Cooper, W. W., & Rhodes, E. (1978). Measuring the efficiency of decision making units. *European Journal of Operational Research*, 2(6), 429–444.
- Coelli, T., Rao, D. S. P., & Battese, G. E. (1998). Review of production economics. In T. Coelli, D. S. P. Rao, & G. E. Battese (Eds.), *An introduction to efficiency and productivity analysis* (eds) (pp. 11–37). Boston, MA: Springer US.
- Coelli, T. J., Prasada Rao, D. S., O'Donnell, C. J., & Battese, G. E. (2005). *An introduction to efficiency and productivity analysis* eds.. Boston, MA: Springer US.
- Cordero, J. M., Santín, D., & Sicilia, G. (2015). Testing the accuracy of DEA estimates under endogeneity through a Monte Carlo simulation. *European Journal of Operational Research*, 244(2), 511–518.
- Cummins, J. D., Weiss, M. A., & Zi, H. (1999). Organizational form and efficiency: The coexistence of stock and mutual property-liability insurers. *Management Science*, 45(9), 1254–1269.
- Dellnitz, A., Kleine, A., & Rödder, W. (2018). CCR or BCC: What if we are in the wrong model? *Journal of Business Economics*, 88(7), 831–850.
- Färe, R., Grosskopf, S., Noh, D. W., & Weber, W. (2005). Characteristics of a polluting technology: Theory and practice. *Journal of Econometrics*, 126(2), 469–492.
- Golany, B., & Storbeck, J. E. (1999). A data envelopment analysis of the operational efficiency of bank branches. *INFORMS Journal on Applied Analytics*, 29(3), 14–26.
- Greene, W. H. (2008). *The econometric approach to efficiency analysis. The measurement of productive efficiency and productivity change*. New York: Oxford University Press.
- Hazewinkel, M. (1992). *Encyclopaedia of mathematics*. Springer Netherlands, Dordrecht.
- Holland, D., & Lee, S. (2002). Impacts of random noise and specification on estimates of capacity derived from data envelopment analysis. *European Journal of Operational Research*, 137(1), 10–21.
- Kaffash, S., Azizi, R., Huang, Y., & Zhu, J. (2020). A survey of data envelopment analysis applications in the insurance industry 1993–2018. *European Journal of Operational Research*, 284(3), 801–813.
- Kohl, S., & Brunner, J. O. (2020). Benchmarking the benchmarks – Comparing the accuracy of data envelopment analysis models in constant returns to scale settings. *European Journal of Operational Research*, 285(3), 1042–1057.
- Kohl, S., Schoenfelder, J., Fügener, A., & Brunner, J. O. (2019). The use of data envelopment analysis (DEA) in healthcare with a focus on hospitals. *Health Care Management Science*, 22(2), 245–286.
- Krüger, J. J. (2012). A Monte Carlo study of old and new frontier methods for efficiency measurement. *European Journal of Operational Research*, 222(1), 137–148.
- Lee, H., Park, Y., & Choi, H. (2009). Comparative evaluation of performance of national R&D programs with heterogeneous objectives: A DEA approach. *European Journal of Operational Research*, 196(3), 847–855.
- López, F. J., Ho, J. C., & Ruiz-Torres, A. J. (2016). A computational analysis of the impact of correlation and data translation on DEA efficiency scores. *Journal of Industrial and Production Engineering*, 33(3), 192–204.
- Mahmoudi, R., Emrouznejad, A., Shetab-Boushehri, S.-N., & Hejazi, S. R. (2020). The origins, development and future directions of data envelopment analysis approach in transportation systems. *Socio-Economic Planning Sciences*, 69, Article 100672.
- Pedraja-Chaparro, F., Salinas-Jimenez, J., & Smith, P. (1997). On the Role of Weight Restrictions in Data Envelopment Analysis. *Journal of Productivity Analysis*, 8(2), 215–230.
- Pedraja-Chaparro, F., Salinas-Jiménez, J., & Smith, P. (1999). On the quality of the data envelopment analysis model. *Journal of the Operational Research Society*, 50(6), 636–644.
- Perelman, S., & Santín, D. (2009). How to generate regularly behaved production data? A Monte Carlo experimentation on DEA scale efficiency measurement. *European Journal of Operational Research*, 199(1), 303–310.
- Resti, A. (2000). Efficiency measurement for multi-product industries: A comparison of classic and recent techniques based on simulated data. *European Journal of Operational Research*, 121(3), 559–578.
- Ruggiero, J. (2005). Impact assessment of input omission on DEA. *International Journal of Information Technology & Decision Making*, 4(03), 359–368.
- Santín, D., & Sicilia, G. (2017). Dealing with endogeneity in data envelopment analysis applications. *Expert Systems with Applications*, 68, 173–184.
- Siciliani, L. (2006). Estimating technical efficiency in the hospital sector with panel data. *Applied Health Economics and Health Policy*, 5(2), 99–116.
- Simar, L., & Wilson, P. W. (2002). Non-parametric tests of returns to scale. *European Journal of Operational Research*, 139(1), 115–132.
- Simar, L., & Wilson, P. W. (2015). Statistical approaches for non-parametric frontier models: a guided tour. *International Statistical Review*, 83(1), 77–110.
- Thompson, R. G., Singleton, F. D., Thrall, R. M., & Smith, B. A. (1986). Comparative site evaluations for locating a high-energy physics lab in Texas. *INFORMS Journal on Applied Analytics*, 16(6), 35–49.
- Tone, K. (2001). A slacks-based measure of efficiency in data envelopment analysis. *European Journal of Operational Research*, 130(3), 498–509.
- van Biesebroeck, J. (2007). Robustness of productivity estimates. *The Journal of Industrial Economics*, 55(3), 529–569.
- Weisberg, H. (1992). *Central tendency and variability* (Thousand Oaks, California).