University of Augsburg
Faculty of Business and Economics
Health Care Operations/Health Information Management

Universität Augsburg
University

Advancing Efficiency Analysis using Data Envelopment Analysis: The Case of German Health Care and Higher Education Sectors

Kumulative Dissertation
der Wirtschaftswissenschaftlichen Fakultät
der Universität Augsburg
zur Erlangung des akademischen Grades eines
Doktors der Wirtschaftswissenschaften
(Dr. rer. pol.)

by
Mansour Zarrin

December, 2021

Erstgutachter:	Prof. Dr. Jens O. Brunner
Zweitgutachter:	Prof. Dr. Robert Klein
Drittgutachter:	Prof. Dr. Yarema Okhrin
Vorsitzender der mündlichen Prüfung:	Prof. Dr. Marco C. Meier
Tag der mündlichen Prüfung:	03.06.2022

List of Contributions

This dissertation contains the following contributions submitted to or published in scientific journals. The specified categories relate to the journal ranking VHB-JOURQUAL 3 of the Verband der Hochschullehrer für Betriebswirtschaft e.V. (2015). The order of the contributions corresponds to the order of print in this thesis.

1. Zarrin, M., Brunner, J.O. (2022). Analyzing the Accuracy of Variable Returns to Scale Data Envelopment Analysis Models
   Status: Submitted to European Journal of Operational Research (Under Review), Category A.

2. Zarrin, M., Schoenfelder, J., Brunner, J.O. (2022). Homogeneity and Best Practice Analyses in Hospital Performance Management: An Analytical Framework. *Health Care Management Science*, DOI: 10.1007/s10729-022-09590-8
   The printed version is a pre-print of an article published in Health Care Management Science. The final authenticated version is available online at: https://doi.org/10.1007/s10729-022-09590-8
   Status: Published in Health Care Management Science, Category A.

3. Zarrin, M. (2022). A Mixed-Integer Slacks-Based-Measure Data Envelopment Analysis for Classifying Inputs and Outputs of German University Hospitals
   Status: Submitted to Health Care Management Science *(Under Review), Category A.*

4. Otto, J. M., Zarrin, M., Wilhelm, D., & Brunner, J. O. (2021). Analyzing the relative efficiency of internationalization in the university business model: the case of Germany. *Studies in Higher Education*, 46(5), 938-950. DOI: 10.1080/03075079.2021.1896801
   The printed version is a pre-print of an article published in Studies in Higher Education. The final authenticated version is available online at: https://doi.org/10.1080/03075079.2021.1896801
   Status: Published in Studies in Higher Education, Not categorized.

Acknowledgment

I would like to thank my advisor, Prof. Dr. Jens O. Brunner, who directed me throughout this dissertation. I am also thankful to my colleagues at the Faculty of Business and Economics at the University of Augsburg, particularly those from the Chair of Health Care Operations / Health Information Management who provided insights into this dissertation.

Most importantly, without my wonderful family, none of this could have happened. To my mom and dad, who have provided me with their unstinting support and encouragement every single day. To my gorgeous sisters, my generous brother, and my loveable niece, this is a testament to your unwavering affection and undying support that this dissertation exists.

– Love you –

Table of Contents

List of Abbreviations

An alphabetical list of the most frequently used abbreviations in this dissertation can be found below.

Abbreviation	***Explanation***
ANN	Artificial Neural Network
AR	Assurance Region
BCC	Banker, Charnes, and Cooper
BPM	Best Practice Model
CCR	Charnes, Cooper, and Rhodes
CRS	Constant Returns to Scale
DEA	Data Envelopment Analysis
DGP	Data Generation Process
DMU	Decision Making Unit
DRS	Decreasing Returns to Scale
IRS	Increasing Returns to Scale
MLP	Multilayer Perceptron
PPS	Production Possibility Set
SBM	Slacks-Based Measurement
SOM	Self-Organizing Map
TCM	Transformative Capacity Model
VRS	Variable Returns to Scale

1 Introduction

Higher education and the health care sectors share many similarities. From an academic perspective, both institutions are service providers who have a close affinity in terms of mission, organizational structure, and resources needed. Higher education and health care are both known for their service missions, which attract people who are involved in social causes. Experienced organizational leaders are well-positioned to capitalize on these similarities, especially during this time of increasing demand and limited supply. The situation is especially relevant in Germany where a sizeable proportion of Gross Domestic Product (GDP) is assigned to health and education sectors. In 2019, 11.9% and 6.4% of GDP were respectively allocated to health and education expenditures in Germany (Federal Statistical Office 2021). More importantly, both of these expenditures have been on the rise during the past few years, as shown in Figure 1. Therefore, an instrument that is most suitable for a hospital or a university as a service provider is the optimization of efficiency, cost, and maintaining the quality of service. There has been an increase in the objective evaluation of performance and making of management decisions across all industries. Most administrators initially reacted to this by cutting costs or avoiding circumstances that would likely waste money, however in a while comprehended they had to improve their performance to remain profitable.

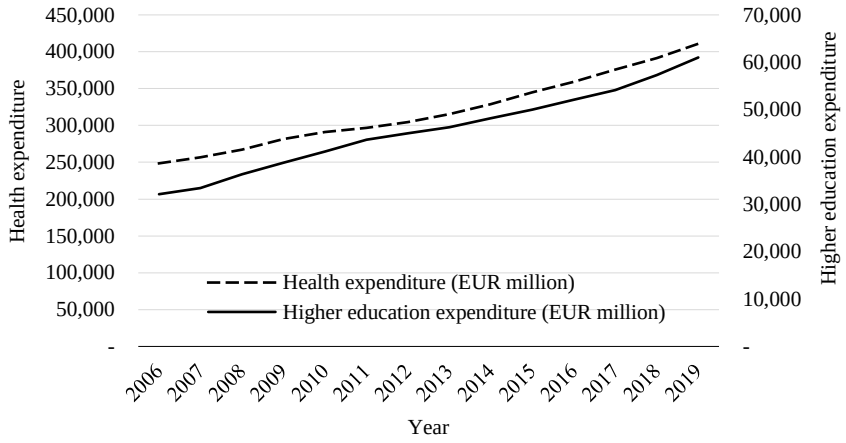


Figure 1. Trends in German health and higher education expenditures (Federal Statistical Office 2021)

Health care and education sectors will continue to face agitated periods and increasing competition. Various global threats, such as the COVID-19 pandemic and climate change, have brought to light the importance of the effective usage of limited resources to deal with these challenges (OECD and Union 2020). To survive, the managers of these institutions must promote and improve performance within their organizations. Improving performance is not a straightforward process. Various aspects of a service, organization, or process should be examined. In some cases, the quality of the service (output) may be improved by increasing resources used (or inputs). While maintaining quality in other cases, more must be accomplished with fewer resources. Organizations need to know their performance based on the employment of resources (inputs) such as materials and labor, as well as the outputs such as

service quality and customer satisfaction. The task of choosing the right balance between the inputs and outputs will always be a challenge for managers in both the health care and education sectors.

In the German health care sector, hospitals account for 40% of health care expenditures, which amounts to over 160 billion euros in 2019 (OECD and Union 2020). Since the introduction of the Diagnosis-related Groups (DRGs) system[1], hospitals in Germany have been under increasing pressure to operate more efficiently. In this context, the pressure on university (or teaching) hospitals to improve their efficiency is even higher, since they are responsible for combining training with patient care under one roof. On the other hand, Germany spent about 105 billion euros on research and development in 2018. Of this amount, 18% was spent by higher education institutions (Federal Statistical Office 2021). Both academics and the general public are paying increasing attention to the efficiency of German universities as a result of tight public budgets (Kempkes and Pohl 2010). A key component of improving universities' efficiency is internationalization, which enhances the quality of their services. A consequence of internationalization is that university missions are altered. In addition to changing funding structures and increased competition for resources among universities at national and global levels, universities also deal with multifaceted environmental factors (Valero and van Reenen 2019). In response to the opportunities and challenges that internationalization presents to universities, institutions of higher education across Western countries have adopted their business models (McAdam et al. 2017).

Measuring efficiency is a central continuous enhancement tool for businesses/services to remain competitive. With performance evaluations, managers can diagnose the areas of strength and weakness of their business operations, activities, and processes. They can also identify opportunities to make difference and assess how to expand new services and processes. Benchmarking was developed as a new technique for evaluating performance. However, benchmarks based on classical analytical schemes (e.g., single-measure-based gap analysis) posed more dilemmas than solutions. For example, ratios such as return on investment and cost per unit can be employed as an indicator of financial performance (Cooper et al. 2007). Nevertheless, they are insufficient to discriminate between *best practices* and to evaluate operational efficiency. The evaluation of performance must not only create benchmarks but also provide information about inefficient organizations and explain how improvement can be achieved. That is what both the education and health care industries need in the present day (Street et al. 2006; Ozcan 2014).

In the literature, there are various methods for performing comparative performance analyses that can be classified as parametric and non-parametric. The least-squares regression and Stochastic Frontier Analysis (SFA) are the most popular parametric frontier approaches. Data Envelopment Analysis (DEA) is a nonparametric approach widely used method of measuring efficiency. In this

[1] DRG aims to standardize payment to hospitals and encourage cost-saving initiatives.

context, *efficiency* is outlined as the amount/number of output units produced through a unit of input used. The production process involves the transformation of inputs (such as labor) into one or more outputs. This process is called the *production function*. The production function, in general, is a quantitative representation of the technology that stipulates the relationship between inputs employed to produce the maximum possible output(s). Based on the input-output vectors that correspond to the observed Decision-Making Unit (DMU), the DEA derives the Production Possibility Set (PPS) (Cooper et al. 2007). A nonparametric approach is taken here which does not make any assumptions about the production function. Therefore, it requires no knowledge of how such transformation occurs. A common misconception is that the DEA approach is only applicable in the absence of input and output prices. To measure cost efficiency, we can substitute the DEA with the SFA when only some regularity settings on the underlying technology can be imposed (Cooper et al. 2007).

Despite the history of efficiency benchmarks starting with the study conducted by Farrell (1957), the theoretical expansion of the basic DEA model (known as CCR) was initiated by Charnes et al. (1978), who defined a measure of efficiency by maximizing the weighted ratio of outputs over inputs for each DMU. Mathematically speaking, by defining input vector $\boldsymbol{X} = (x_{1o}, \dots, x_{mo})$ and output vector $\boldsymbol{Y} = (y_{1o}, \dots, y_{so})$ the overall efficiency of $DMU_o \;\; \forall o \in N = \{1, \dots, n\}$ can be formulated as $TE_o = \sum_{r=1}^{s} u_r y_{ro} / \sum_{i=1}^{m} v_i x_{io}$ where, u_r and v_i are the weights attached to output r and input i, respectively. Corresponding to the terminology of efficiency developed in the literature (Street et al. 2006), TE_o represents *technical efficiency*. Therefore, the mathematical program of the CCR DEA model in ratio form is presented as follows:

$$\max TE_o = \frac{\sum_{r=1}^{s} u_r y_{ro}}{\sum_{i=1}^{m} v_i x_{io}} \tag{1.1}$$

$$\text{s.t.} \quad \frac{\sum_{r=1}^{s} u_r y_{rj}}{\sum_{i=1}^{m} v_i x_{ij}} \leq 1, \;\; \forall j \tag{1.2}$$

$$\boldsymbol{u}, \boldsymbol{v} \geq 0 \tag{1.3}$$

Eq. (1.2) represents the only and important constraint that the CCR DEA model is subjected to: none DMU can get a technical efficiency greater than 1. The weights (u_r and v_i) are the central feature of the CCR DEA model. They are optimized through solving the model to cast the DMU under evaluation in the best possible light so that no other set of weights can yield a greater value than TE_o. The objective function of Model (1) is non-linear. To deal with non-linearity, either the summation of weighted inputs in the numerator or the summation of weighted outputs in the denominator of Eq. (1.1) must equal 1 (Coelli et al. 2005). Then, we can rewrite the model by adding a constraint (summation of weighted inputs equals to 1) and operating Eq. (1.2) as Model (2). This is recognized as the multiplier form of the CCR DEA model. Using this form of the model, the orientation of an optimal production plan can be adjusted. Alternatively, this model can also be articulated as a dual counterpart model

(known as envelopment form), which has the advantage of requiring fewer constraints and is more known than Model (2) (Coelli et al. 2005).

Multiplier CCR DEA			*Envelopment CCR DEA*	
	$\max \sum_{r=1}^{s} u_r y_{ro}$	(2.1)	$\min \theta_o$	(3.1)
s.t.	$\sum_{i=1}^{m} v_i x_{ij} - \sum_{r=1}^{s} u_r y_{rj} \geq 0, \forall j$	(2.2)	s.t. $\sum_{j=1}^{n} x_{ij} \lambda_j \leq \theta_o x_{io}, \forall i$	(3.2)
	$\sum_{i=1}^{m} v_i x_{io} = 1$	(2.3)	$\sum_{j=1}^{n} y_{rj} \lambda_j \geq y_{ro}, \forall r$	(3.3)
	$\boldsymbol{u}, \boldsymbol{v} \geq 0$	(2.4)	$\boldsymbol{\lambda} \geq 0$	(3.4)

Note that in Model (3), θ_o represent the technical efficiency index of DMU_o and λ_j are the intensity variables. In the CCR DEA model, the assumption of Constant Returns to Scale (CRS) is underlined. The PPS is assumed to have the following property: if (x, y) is a feasible point, then $(a \cdot x, a \cdot y)$ is also feasible for any $a > 0$. It is possible to modify this assumption to allow for different postulates for PPS. The CCR model has been extended in various ways since the beginning of DEA studies, and among them, the BCC (Banker-Charnes-Cooper) model is representative (Banker et al. 1984) in which a more flexible Variable Returns to Scale (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 envelopment form of CCR DEA model (Model (3)). When not all DMUs operate at the optimal scale, then this model might be appropriate. It is often the case in the health care sector that inefficient scale is the result of flawed competition, financial constraints, and governing restrictions on mergers (Cooper et al. 2007). Choosing between CRS and VRS is therefore a complex decision that requires a thorough understanding of the limitations of the market in a particular sector. In the case of a hospital operating at a suboptimal scale, if the CRS setting is applied, then the estimates of technical efficiency might be skewed by scale efficiency effects (Street et al. 2006).

Consider Model (2), there might be many zeros in the optimal weights of the model, indicating that the evaluating DMU may have a weakness in the factors (inputs and outputs) compared to efficient DMUs. Furthermore, a significant difference in weights between items may also be a cause for concern. Having no control over the boundaries of optimal weights leads to the emerging Assurance Region (AR) DEA model, which constrains the weight of special inputs/outputs relative to others (Thompson et al. 1986). For instance, one can limit the region of weights to some special area by adding this constraint $lb_{1,2} \leq {}^{u_1}/_{u_2} \leq ub_{1,2}$ to Model (2), where $lb_{1,2}$ and $ub_{1,2}$ designate lower and upper bounds that the ratio of weights for outsputs 1 and 2 may assume.

Both CCR and BCC DEA models are radial where inputs are proportionally reduced and outputs are proportionally expanded. This assumption can sometimes be too 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. Therefore, the radial models

may delude a decision-making process when these slacks have a significant role to play in estimating efficiency. These limitations lead to the expansion of non-radial models. Slacks-based Measure (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 CRS setting is a non-linear model that can be reformulated as a linear counterpart by using Charnes–Cooper transformation approach (Charnes and Cooper 1962) as follows:

SBM DEA

$$\min \tau_o = \frac{1-\frac{1}{m}\sum_{i=1}^{m}\frac{s_i^-}{x_{io}}}{1+\frac{1}{s}\sum_{r=1}^{s}\frac{s_r^+}{y_{ro}}} \quad (4.1)$$

$$\text{s.t.} \quad x_{io} = \sum_{j=1}^{n} x_{ij}\lambda_j + s_i^- , \forall i \quad (4.2)$$

$$y_{ro} = \sum_{j=1}^{n} y_{rj}\lambda_j - s_r^+ , \forall r \quad (4.3)$$

$$\boldsymbol{s}^-, \boldsymbol{s}^+, \boldsymbol{\lambda} \geq \mathbf{0} \quad (4.4)$$

Non-oriented transformed SBM DEA

$$\min \rho_o = t - \frac{1}{m}\sum_{i=1}^{m}\frac{s_i^-}{x_{io}} \quad (5.1)$$

$$\text{s.t.} \quad t + \frac{1}{s}\sum_{r=1}^{s}\frac{s_r^+}{y_{ro}} = 1 \quad (5.2)$$

$$t \cdot x_{io} = \sum_{j=1}^{n} x_{ij}\lambda_j + s_i^- , \forall i \quad (5.3)$$

$$t \cdot y_{ro} = \sum_{j=1}^{n} y_{rj}\lambda_j - s_r^+ , \forall r \quad (5.4)$$

$$\boldsymbol{s}^-, \boldsymbol{s}^+, \boldsymbol{\lambda} \geq \mathbf{0}, t > 0 \quad (5.5)$$

where τ_o and ρ_o are the SBM-efficiency in non-linear and transformed forms and they are units-invariant. In other words, there is no dependence on the units of measurement of inputs or outputs for SBM efficiency values. $\boldsymbol{s}^-$ and $\boldsymbol{s}^+$ 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 Model (5) be $\{\rho^*, t^*, \boldsymbol{\lambda}^*, \boldsymbol{s}^{-*}, \boldsymbol{s}^{+*}\}$. The optimal solution of the SBM DEA model can be defined as $\{\rho^*, t^*, \boldsymbol{\lambda}^*/t^*, \boldsymbol{s}^{-*}/t^*, \boldsymbol{s}^{+*}/t^*\}$. The output (input)-oriented SBM model can be derived from Model (4) by abandoning the numerator (denominator) of the objective function of the non-oriented SBM DEA.

Now let us examine the production functions of radial and non-radial DEA models with an example of five DMUs that use two inputs (x_1 and x_2) to produce one unity output (y) (see Figure 2). To draw the DEA isoquants, inputs must be rescaled so that each input is divided by the output level which is equal to one. DMUs A, B, and C form the efficient frontier production, whereas D and E are inefficient. The PPS is built on the area enclosed by the efficient frontier (solid line) plus the horizontal line prolonging to the right from C and the vertical line ascending from A (dashed lines). CCR- and SBM-efficiency scores for inefficient DMUs (D and E) are presented by θ^* and ρ^*, respectively. The role of nonzero slacks in the CCR model is illustrated by DMU E. The radial projection of E (E'_{CCR}) encounters the virtual frontier going up from A where is not naturally enveloped. However, the E'_{CCR} is not efficient. One could still reduce the amount of x_2 by $\overline{E'_{CCR}A} = 2$. The difference between the two points E'_{CCR} on the virtual frontier and A on the edge of the frontier indicates the slack for x_2. However, the SBM efficiency reflects nonzero input/output slacks when they exist. Consider the inefficient DMU E, SBM model projects this DMU along the efficient frontier (gray arrows) onto the DMU A ($E'_{SBM} =$

A). This results in $s_1^{-*} = 0.5$ and $s_2^{-*} = 4.0$ for DMU E. Similar investigation can also be conducted on inefficient DMU D.

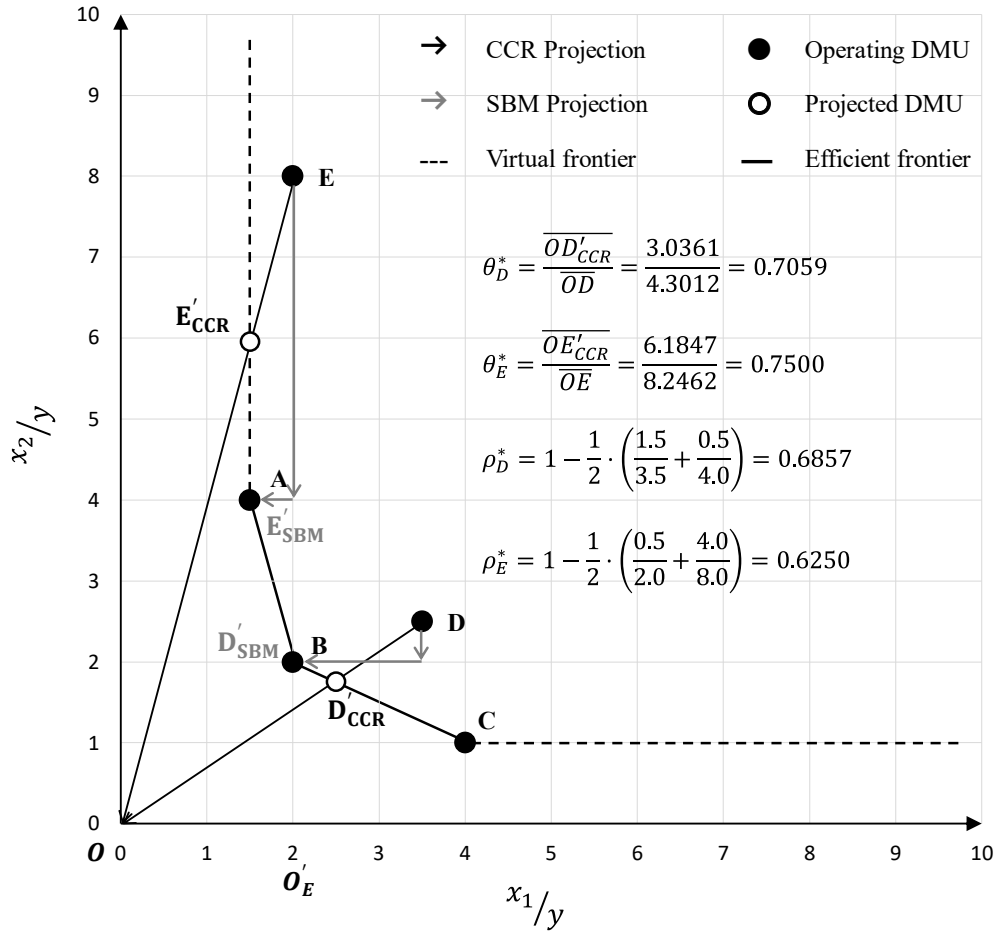


Figure 2. An illustration of radial and non-radial efficiency assessment

DEA modeling comes in many forms (see, for example, Cooper et al. (2007) and Tone (2017)), and the models presented here are merely a few of those diverse forms. However, it has been well developed and documented in the literature from both theoretical and practical viewpoints. Over the last ten years, this trend has become exponentially more prevalent. Emrouznejad and Yang (2018) report that over 10,000 DEA-related articles have been published to date. Among the existing applications of DEA models, measuring the efficiency of health care and education (including higher education) are notable in the early days of DEA as mentioned by Liu et al. (2013). During the early stages of DEA development, education attracted the most attention. This may be due to the study of public education efficiency conducted by Charnes et al. (1981). The vast majority of publications in the health care area concern hospitals (Kohl et al. 2019). Some of the other applications are for nursing homes, physician practice and disease-specific, home health agencies, and other health care organizations (Ozcan 2014). Practitioners pay very close attention to the application and use of DEA as its popularity grows. However, there are several procedural issues in the application of DEA that should be addressed such as those that pertain to the homogeneity of the DMUs (Dyson et al. 2001). In practice, these issues can

pose difficulties. Additionally, it is important to handle data irregularities and issues related to structural complexity. These include accommodating flexible measures that can be designated as either input or output as well as conditions in which one or more inputs/outputs can take only integer quantities (Zhu and Cook 2007). For a study of hospital efficiency, consider the number of medical interns or the number of graduate students at a university. In addition to their role as output measure, these factors can serve as input since they are a key component of the organization's total staff complement.

The main goal of this dissertation is to investigate the advancement of efficiency analysis through DEA. This is practically followed by the case of German health care and higher education organizations. Towards achieving the goal, this dissertation is driven by the following research questions:

1. How the quality of the different DEA models can be evaluated?
2. How can hospitals' efficiency be reliably measured in light of the pitfalls of DEA applications?
3. In measuring teaching hospital efficiency, what should be considered?
4. At the crossroads of internationalization, how can we analyze university efficiency?

The remainder of this dissertation is organized as follows. Section 2 summarizes all contributions included in this dissertation. Note that the appendix provides a complete copy of each contribution. In Section 3, the research questions are addressed in detail. Finally, this dissertation comes to a close with some conclusions in Section 4.

2 Summary of Contributions

A summary of the contributions of this dissertation to the current literature is presented in this section. In the appendix, you will find a complete version of each contribution.

2.1 *Analyzing the Accuracy of Variable Returns to Scale Data Envelopment Analysis Models*

As mentioned earlier, a critical component of decision-making management is evaluating efficiency to reduce resource waste and identify better performers. The most representative of the non-parametric approaches developed for efficiency analysis is DEA. Though DEA development and applications have progressed substantially in the last five decades, there remains no superior DEA model. As a matter of fact, the basic models (CCR and BCC) continue to be dominant in various applications such as health care despite known issues such as remaining slacks and zero weights. In Contribution 1, we mainly focus on VRS settings as the most commonly employed RTS in the DEA literature, and we try to determine if the prominence of the BCC model in the literature is justified. Without an appropriate benchmark to compare various DEA models, the development of a *gold standard* is not likely to be possible. Because of this, DEA is still viewed primarily as a scientific topic rather than an operational tool. Contribution 1 fills this gap by making DEA models comparable in accuracy based on VRS settings. It is impossible to evaluate the quality of the DEA estimates since *true efficiency* cannot be determined directly from the empirical data. It is however possible to generate artificial datasets using Monte Carlo simulations under certain assumptions and regimes to address this issue.

Earlier studies in the Data Generation Process (DGP) have primarily used the Cobb-Douglas (CD) production function. This is because of the complications of the alternatives, including simple and convex constraints, that are imposed by microeconomic regularity conditions. Researchers such as Perelman and Santín (2009) have shown that CD is limited in imposing input substitution elasticity at one level and fixed-scale economies. In order to generate more testable production data, Translog (transcendental logarithmic) has emerged as a generalized form of CD. Contribution 1 provides then a way for the Translog production function to be used under pure VRS settings so as to extend its applicability. A majority of DGP studies use only one adjustment for input numbers. In general, scenario generation has not received enough attention. Contribution 1 also confirms that most studies vary three or fewer characteristics in the used DGP. CCR and BCC models have largely been studied beforehand, and their properties have sometimes been compared to parametric approaches. Nevertheless, the alternatives to these models (such as SBM and AR) are sparse. The robustness of the previous studies is also a concern. Based on the random data used in DEA estimations, it is impracticable to dispute the replication for each scenario. Contribution 1 addresses these issues by providing a process allowing us the comparison of the accuracy of DEA models under the VRS assumption. Contribution 1 demonstrates, through questions regarding the reliability of DEA results and by analyzing the efficiency

assessment process, that the environment of DEA applications has a significant impact on their accuracy.

By advancing the scenario variation significantly, Contribution 1 aims to improve the general result validity. An important component of studying the quality of DEA models is designing a sophisticated DGP to produce well-behaved data for the DMUs, as mentioned earlier. In order to compare the estimated efficiencies from different DEA models with their true efficiency scores, we then generate artificial data to identify the true efficiency of each DMU. As a result, Contribution 1 advances the scenario variation significantly to increase the general result validity. The generated scenarios are concrete arrangements of varying values of all characteristics of the DGP (e.g., number of inputs and DMUs, and importance of inputs). In contrast to the literature, which has used no more than 1300 scenarios, Contribution 1 analyzes 7,776 different scenarios. It also examines the coverage of different characteristics of a DEA study in addition to the number of scenarios in an attempt to evaluate whether the environment of the study influences the accuracy of the results.

By using ten distinct characteristics with varied levels Contribution 1 advances the literature. As shown in Contribution 1, the monotonicity and curvature requirements are directly imposed by developing a mathematical model. As a result, valid scenarios are generated with VRS properties. As a result of the methodology proposed, a more sensible DGP can be guaranteed. By decomposing input substitution into substitutability and substitution distribution, realistic and well-behaved DMUs are also guaranteed. In terms of robustness, it is found that the number of DMUs is highly correlated with the number of replications in each scenario. The moving standard deviation of the benchmark value must therefore serve as an elastic stopping condition for replications of each scenario. In Contribution 1, the BCC model is compared with two other DEA models: AR and SBM. The results are compared using two methods: multiple performance indicator benchmark scores and DEA-based hypotheses testing. There are some important properties of an efficiency estimator introduced in the literature that are covered by the benchmark score. Contribution 1 addresses the statistical properties of DEA's estimators by comparing estimations with the true efficiencies as a means of establishing a statistical foundation for DEA.

Analyzing the accuracy of VRS DEA models indicates that the AR and SBM models both perform considerably better than the BCC model, which is most commonly used in DEA applications. The BCC model performs better than CCR among basic DEA models, which is not a surprise since DGP is built on VRS settings. Nonetheless, it can be taken as a positive sign that the DGP results are reliable and its working mechanism is clearer. The main conclusions drawn from analyzing the performance indicators are also reinforced by the results of hypothesis statistical tests. Compared with basic DEA models, the number of rejected scenarios in AR and SBM models is much lower. In addition, the results of these tests indicate the importance of selecting the right RTS, as, on average, the CCR DEA model fails to estimate the efficiency scores of 50% of scenarios generated under the VRS setting.

In this case, also, the AR and SBM models are undoubtedly outperforming the BCC model. Contribution 1 also confirms that the use of more inputs and a low number of DMUs both negatively affect accuracy. Furthermore, they lead to more rejected scenarios. By decreasing the lower bounds of true efficiencies, the accuracy drops slightly and the rejected scenarios increase trivially. Therefore, allocating a greater share of DMUs to the true efficiency frontier marginally reduces the quality of DEA models. When input importance is different for all inputs, the accuracy of all DEA models is to some degree lower than when all inputs are equally important. In terms of the results of the substitution distribution of inputs, all DEA models perform better when the input substitution is unequal. In the basic DEA models CCR and BCC, the performance is significantly diminished by increasing the scale effect value. In general, the AR and SBM models do better when the extent of scale effects is raised to increase the curvature of the production function. Across all models, it appears that larger input ranges lead to worse results. Finally, Contribution 1 concludes the prominent positioning of the AR model without a special tuning of the virtual weight restrictions. For some applications, however, establishing weight bounds may be too complex for explicit articulation. Based on the fact that the SBM model performs almost as well as the AR model, it is recommended that the SBM model be adopted as the VRS DEA standard in which weights are not pre-supposed.

2.2 *Homogeneity and Best Practice Analyses in Hospital Performance Management: An Analytical Framework*

The literature has steadily gained attention to hospital performance modeling using DEA models, as discussed earlier. Within the standard DEA framework, DMUs are frequently assumed to be functionally similar and therefore homogeneous. Therefore, hospitals' inefficiency is supposedly caused by inefficient input use to create outputs. However, the difference in efficiency scores might be caused by non-homogeneous DMUs. Despite being used a lot as a benchmarking tool, the traditional DEA framework lacks predictive capabilities. In practice, predicting feasible levels of performance is a critical step towards achieving better performance than competitors, especially in the face of limited resources. Therefore, Contribution 2 aims to tackle these two common issues by developing and evaluating a framework for analyzing the performance of a large set of German hospitals.

The homogeneity assumption is normally applicable in a sample based on implicit knowledge of the DEA investigator. Dyson et al. (2001) describe three major homogeneity assumptions made by DEA with regards to the DMUs under evaluation. They are resource similarity, functional similarity, and environment similarity. A DEA application can be heavily influenced by ignoring either of these assumptions. The DEA structure can be used to model explicitly these differences by identifying, defining, measuring, and then modeling them. However, even if all influential environmental variables can be considered, there will be less discrimination since this will lead to a substantial increase in inputs and outputs. The managers of inefficient hospitals usually request further information after being

presented with the results of the DEA such as keep close track of progress by analyzing what-if scenarios and setting performance goals during operational phases. Thus, hospitals must be able to set up actionable and specific performance targets. In spite of successful modeling to measure comparative efficiency among competing units, little attention has been given to integrating predictability into hospital performance measurement frameworks.

In this contribution, we introduce a three-stage approach to analyzing market homogeneity and providing predictive capabilities to the DEA. By grouping similar DMUs based on their transformation capacity (or technology), environmental variables can be modeled implicitly. In Stage 1, an artificial neural network architecture based on a self-organizing map (SOM) is developed to cluster hospitals based on their similarity in terms of transformation capacity. In this way, categories are discovered in the multidimensional and large dataset of DMUs. According to the environment of the DMUs, the hospitals are clustered into homogeneous sets. Specifying the appropriate number of clusters is an important issue in clustering. The quality of partition and cluster validity has been assessed by several authors using different indices. Contribution 2 calculates three well-known criteria called the Caliński-Harabasz, Silhouettes, and Davies-Bouldin to assess the quality of hospital clusters obtained from various configurations defined for the SOM networks. In addition, we cluster hospitals based on their natural clustering characteristics, which are typically the size (number of beds) and ownership type. By comparing the quality indicators calculated for SOM clustering with those calculated for natural clustering, we show that natural clustering cannot produce high-quality clusters for hospitals. Because of this, they cannot ensure homogeneity within clusters.

Using an input-oriented SBM-DEA model, we calculate the efficiency score and projection of each hospital in each cluster in Stage 2. These estimates are then tested versus bootstrapped DEA estimates to determine whether or not they are biased upward. We also conduct DEA-based hypotheses tests to compare two groups of hospitals. We perform further appropriate tests after indicating the existence of a statistical difference between two clusters of hospitals to determine which group's efficiency distribution is stochastically greater than the others. Using this method, we identify each hospital cluster's leader and follower. To address heterogeneity, we analyze two aspects of clusters: transformative capacity and scale heterogeneity (scalability). Identified leaders and followers play a major role in this process. An MLP-ANN architecture is trained for each cluster to learn the relationship between inputs (input layer) and outputs (output layer), i.e., transformative capacity model (TCM). Next, the trained networks are used to simulate outputs for each cluster using the TCMs from all other clusters. It resembles the transformative capacity of a given cluster by which the set of inputs is transformed into the outputs. Next, we employ the comparison procedure developed in Stage 2 to investigate whether using the simulated outputs generated by the TCM of the leader increases the efficiency distribution of hospitals underlying in the following cluster. If it improves, then there is a reason to suggest the disparity between the (original) SBM-DEA efficiencies of the leading and

following clusters can be partially explained by differences in transformational capacity. To analyze scale heterogeneity (scalability), the simulated outputs of a leader obtained from the TCM of its follower are used to recalculate the efficiency of the leader. We then perform the comparison procedure developed in Stage 2 to determine whether the identified leaders (followers) remain as leaders. As long as a leader remains a leader, it is plausible that a part of the disparity between the efficiency distributions of Followers and Leaders stems from scale heterogeneity. Leaders can still achieve greater relative efficiency despite their less efficient transformational capacity process. We also develop a new MLP-ANN architecture in the second part of Stage 3 to predict the best performance level beyond the indirect measure of efficiency scores. Each of these trained networks is called the best practice model (BPM). Here, we use both the inputs and outputs of the hospitals as inputs and the bootstrapped SBM-DEA efficiency scores as outputs node for the MLPs. An estimate of the relative efficiency score of the projected hospital is made using the efficient patterns learned by BPM.

The framework analyzes hospital data compiled by the Federal Joint Committee of Germany in 2017. As the dataset is vast and complex, many preprocessing steps are involved in each stage of the process to ensure accuracy and robustness in calculations. The framework can assist decision-makers by identifying improvement and what-if scenarios. Environmental variables can be accounted for without adding additional variables to DEA models to address non-homogeneity. The results show that clustering hospitals according to ownership or size does not show heterogeneity within groups of hospitals, nor does it reveal homogeneity among groups of hospitals. Further, it is shown that the distribution underlying bootstrapped DEA estimates is not different from that underlying SBM estimates. The SBM estimates of DEA are therefore not significantly skewed upwards. Different levels of efficiency in some German hospitals can be attributed to differences in their transformation capacities and scale heterogeneity rather than inefficient input usage. It is finally shown that training the BPMs to replicate the nonlinear mapping and predictive abilities of DEA models compensates for the lack of predictability of the DEA models.

2.3 *A Mixed-Integer Slacks-Based-Measure Data Envelopment Analysis for Classifying Inputs and Outputs of German University Hospitals*

Aside from their reputation as reliable methodologies, DEA models are also seeing rapid expansion in their use, especially in public sectors such as health care and higher education. Contribution 3 analyzes the performance evaluation of university hospitals from a practical standpoint. Teaching hospitals provide both patient care and medical education, so they are more expensive than their non-teaching counterparts (e.g., acute and general hospitals). Teaching and research as an academic mission should be captured appropriately by defining appropriate measures in the performance assessment process. Almost all past studies have used the basic DEA models to evaluate teaching hospital performance and fail to acknowledge two significant challenges that exist in real-world situations: integer-valued

amounts and flexible measures. In Contribution 3, these two issues are adequately discussed after studying some recent publications on the performance assessment of teaching hospitals.

DEA models traditionally use real (continuous) values as inputs and outputs. There are, however, many situations in which one or some of the inputs/outputs are unavoidably integer values. An example would be the number of beds (input) and outpatients (output) of a hospital. After identifying the inputs and outputs of a DEA application, the first step is determining the suitable technology or PPS. Assumptions about feasible operating points of CRS and VRS are convex combinations of evaluation units, but neither of them considers the integrality constraints of some inputs or outputs. The integration constraints imposed by rounding off the integer values may have a significant effect on the optimality of the solution. When integer inputs and outputs are treated as real values, rounding up (or down) of them arbitrarily may cause infeasibility (outside of the PPS). Moreover, in the standard DEA model, measures (factors) are classified as either inputs or outputs. However, in certain situations, some measures can play either an input or an output role. Consider, for instance, the number of interns or nurses in a teaching (university) hospital. Measures such as these can be either input (two human resources available to the hospital) or output (experienced staff, thus allowing the hospital to take advantage of teaching/research funds). DEA literature refers to these measures as flexible measures.

Contribution 3 presents an SBM model that combines flexible and integer measures simultaneously. All inputs, outputs, and flexible measures in this model can take both real and integer quantities without fluctuating efficiency levels. In addition, the proposed SBM model directly calculates the technical efficiency score, and inflation of scores is prevented by modifying input and output inefficiencies. Using the MILP approach, the proposed model can be solved by most non-commercial and open-source solvers. Additionally, slack values for inputs, outputs, and flexible measures derived from the proposed model are reported and compared with the literature. As a practical matter, Contribution 3 uses a real dataset from 28 German public university hospitals in 2017. Data were obtained through a variety of research methods, including visiting the websites of the hospitals and contacting the departments directly (e-mail/telephone inquiries). This proved very time-consuming and difficult. Inputs include the number of beds, physicians, and nurses. Beds are an integer input measure. Physicians and nurses, on the other hand, are measured in full-time equivalent (FTE) units, i.e., real values. As outputs, the number of outpatients and case-mix adjusted discharges for inpatients are depicted as integers and reals, respectively. Despite their importance, these two major hospital outputs do not include teaching functions. Accordingly, the number of students is used as the integer value of the output of the university hospitals. Additionally, contribution 3 introduces two more flexible measures for assessing the teaching function: graduates and third-party funding income. Graduates can play one of two roles at a university hospital: an input (a resource available to faculties) or an output (experienced staff who benefit from teaching funding). In the efficiency evaluation of university hospitals, third-party funding income can be interpreted similarly; as input (a form of income received)

or as output, as most research-granting agencies will allocate funds to university hospitals with the greatest impact.

This contribution uses a sample of university hospitals with an average of 1,475 beds. Physicians and nurses are employed by them at more than 25,000 and 34,000 FTE, respectively. In terms of output, 2.8 million adjusted inpatient admissions and 11.4 million outpatient visits are recorded. There are approximately 11 thousand graduates and 84 thousand medical students in these units where they have received over €1.5 billion in research funding. Results indicate the 7 university hospitals are characterized as efficient DMUs with the optimum slacks of zero. Most university hospitals treat "Third-party funding income" and "Graduates" as outputs in the final PPS. Moreover, Contribution 3 shows that the PPS is not comparable in certain situations where flexible measures can have a significant impact. Convex PPS slacks (generated by the non-integer DEA) are usually real-valued amounts and the integer output slacks reported by the models are not always a rounding up or down of real-valued slacks. In addition, using the optimal solutions obtained from the models, the inefficiency scores for each inefficient university hospital are decomposed in order to analyze their magnitudes and causes. By far, most of the sources identified by the model are input inefficiencies. Through an examination of magnitudes and sources of inefficiency, this decomposition can provide managers or policy-makers with enlightening information about how to become an efficient university hospital.

2.4 *Analyzing the Relative Efficiency of Internationalization in the University Business Model: The Case of Germany*

There has been a significant increase in competition/market-driven behavior in the German higher education sector, leading to the development of new business models. An important component of these models is the internationalization of university services. Universities are impacted by internationalization and their missions are changed as a result. Due to the opportunities and challenges internationalization poses to universities, many universities adopted business model approaches to respond. Increasing scrutiny of universities by the public and policymakers has led to a growing interest in exploring how universities can use public resources effectively to accomplish their institutional missions. The performance of a business model is determined by two measures: effectiveness (doing the right things) and efficiency (doing things right). The latter is the subject of this contribution, which examines internationalization in university missions in relation to university business models. To do this, Contribution 4 calculates the relative efficiency of the universities in terms of their overall and internationalization relative efficiency. To investigate the relative internationalization and overall efficiency of German universities, Contribution 4 develops a three-stage approach based on outlier detection, SBM DEA models, and regression/correlation analyses to evaluate the effect of environmental variables. As a result, we would be able to learn if universities are effective in their

pursuit of internationalization. Moreover, this enables us to explore the relationship between relative internationalization efficiency and the overall relative efficiency of German universities.

The total annual expenditures and the total number of academic staff are the inputs used to evaluate overall efficiency. The inputs come from human/financial resources invested by the government and other institutions. The traditional university missions of teaching, research, and service should be considered when evaluating outputs for measuring overall efficiency. Total graduates represent teaching. Citations represent the quality of publications produced by researchers. Patenting, as represented by the total number of patent filings, represents service to society through knowledge transfer. As inputs to internationalization, we use total international staff, total funding from the EU and other international organizations. The university internationalization outputs are defined as what is produced by the inputs. They are total international professors, full-time international students, incoming students from the EU's ERASMUS exchange program, and outgoing students from ERASMUS. The last output at this level reflects administrative efforts allocated to the diverse international objectives universities pursue. As an indicator of environmental influences, the total area, total population, and gross domestic product per capita of university municipalities are used in regression and correlation analyses. A drawback of conventional DEA models is their sensitivity to outliers (Dyson et al. 2001). Therefore, in the first stage (pre-processing), this study uses a super-efficient DEA model to identify and exclude outliers. In the second stage, processing, an input-oriented SBM DEA model is used to estimate relative efficiencies for internationalization and overall performance. As a post-processing process, regression and correlation analyses are used in the last stage to evaluate the impact of environmental factors on the efficiency scores.

One university was identified as an outlier by the outlier detection. After excluding this unit from our sample, we move on to the next stage, which includes the rest of the universities. From the results of the second stage, only eight universities were found to be relatively efficient in both overall efficiency and internationalization respects and most universities were regarded as relatively efficient in only one respect. It is interesting to note that all university sizes appear in the efficient frontier, indicating that size is not a predictive factor. Compared with other studies in the literature on mission accomplishment relative efficiency in universities in Germany, these results are confirmed as accurate. Moreover, according to the correlation analysis of efficiency scores, the relationship is statistically significant, but not particularly strong. The contextual variables for university location do not show a strong relationship between variations at either level of analysis. Thus, the relative efficiency of other factors in estimating university efficiency in fulfilling its mission may also be ineffective, which suggests a better understanding of this issue may only be possible when all factors are considered together. This contribution suggests that the relative efficiency of different components may not be able to estimate the relative efficiency of overall university performance in meeting missions. The whole picture can only be understood when all components are considered together.

3 Discussion of Contributions

This section discusses the four research questions presented in the dissertation's introduction. Each of the contributions summarized in Section 2 fills a research gap in the existing literature. There are four sections in which the contributions are discussed in terms of how they provide answers to each research question. These questions are addressed through the integration of the findings of each contribution.

3.1 *Question 1: How the quality of the different DEA models can be evaluated?*

There is no denying that the CCR and BCC models are the most widely used. It may seem natural at first since the DEA has been represented by CCR and BCC models. It is impressive, however, when you consider how much effort has gone into developing models over the years. Despite the fact that the DEA model is extensively used for efficiency estimation, not enough papers try to determine the DEA model(s), which can give the most precise efficiency estimation. Bringing into question the reliability of DEA results when its environment changes, is the primary motivation for a study on the quality assessment of DEA models. RTS settings for DEA applications are most commonly defined as VRS. Possibly, this is a result of the fact that production technology may exhibit decreasing, constant, and increasing returns to scale. Thus, it is essential to determine whether the predominant position of the BCC model in the DEA applications is justified. Since true efficiency is not known in the empirical data, it is impossible to evaluate the quality of the DEA estimates. It is necessary to generate artificial data that mimic the behavior of production function in order to identify the true efficiency of each DMU. This enables us to compare estimations derived from different DEA models with the equivalent true efficiency scores. By doing so, we can make statements about the DEA models' quality and answer the question. Generalized production functions are crucial here since they allow more testable production data to be generated. According to previous studies, Cobb-Douglas functions are used primarily to simulate artificial data. This can be explained by the complexities of the alternatives, such as the monotonicity and convexity constraints imposed by microeconomic behavioral regularities. Cobb-Douglas is limited in imposing the substitution elasticity of one and fixed-scale economies. Therefore, the Translog production function emerges as a generalization of the Cobb-Douglas function.

Another aspect to consider is reflecting different real-world situations that can occur when DEA models are used to measure efficiency. In this way, we can see how accurate DEA models are performing under various conditions. The number of DMUs and the total number of inputs and outputs can, for example, be adjusted to model several real situations. In order to achieve the general validity of the quality results, scenario variation should be significantly increased. Each generated scenario must represent a concrete arrangement of varying values of all the characteristics taken into account during the artificial data generation process. There have been no studies using more than 1300 scenarios up to now, based on the literature. As well as the number of scenarios, different characteristics with diverse

levels should be included to determine if the environment in which a DEA study is conducted affects the accuracy of the results. Answering this question requires also understanding that the DEA's estimations are the basis for any quality judgment. A DEA statement on the quality of models must address four major areas: identifying inefficient DMUs, ranking DMU efficiency, improving the efficiency, and evaluating the overall efficiency of units. Statistical properties of DEA estimators should also be included in the statements when assessing the quality of DEA models. This problem can be addressed by performing statistical tests to compare the distribution of estimates with the actual efficiencies.

3.2 *Question 2: How can hospitals' efficiency be reliably measured in light of the pitfalls of DEA applications?*

DEA has some limitations despite its power when it comes to assessing hospital performance. Three basic issues that DEA investigators should address are selecting appropriate inputs and outputs, selecting the right RTS, and orienting the DEA model. One pitfall in selecting inputs and outputs is to include too many factors. Due to the flexibility of DEA in choosing the weights on the inputs/outputs, the more factors involved, the lower the level of discrimination. Therefore, being sparing with the factors can increase discrimination. When it comes to the resources (inputs), if you can price them, you can eliminate flexible weights and replace them with fixed prices. Discrimination can be reinforced on the output side by removing performance measures that do not closely relate to hospital goals. Hospitals may be too small to operate efficiently or too large to manage. CRS DEA models do not accommodate such situations. As an alternative, VRS models have been designed specifically to account for scale effects during analysis. Regardless of whether variable returns to scale exist, the VRS models encompass the data more closely than the CRS model. A VRS model, where there are no inherent scale effects, tends to overestimate the efficiency of both large and small hospitals. When it is unknown whether the production technology of hospitals exhibits VRS or CRS, hypothesis tests must be conducted to determine the scale effects.

In addition to these issues, the application of DEA also poses a number of issues related to the homogeneity of the hospitals and their environment. DEA makes certain assumptions about the homogeneity of the hospitals being assessed. Hospitals are naturally perceived as being similar in many ways. In order to create a common set of outputs for hospitals in the sample, the hospitals are assumed to perform similar activities and provide comparable services. The second assumption is that all hospitals have access to a similar range of resources. This could include personnel, raw materials, and equipment. As long as a common denominator is established, such as the price of the equipment, then comparisons can still be made, even if different equipment is being used. There is also the tacit assumption that hospitals serve in comparable environments, considering that environmental factors have a significant impact on overall hospital performance. This assumption can rarely be made with

confidence, and therefore, environmental variables are often incorporated into the analysis as a means of supplementing the input/output set.

A basic pitfall in measuring the efficiency of hospitals arises from simply attempting to compare non-homogeneous hospitals. There are, for instance, non-teaching hospitals and teaching hospitals. In this case, the pitfall resides in the fact that teaching hospitals have inevitably more expensive delivery processes than general hospitals. In teaching hospitals, teaching and research are essential functions, so any normal assessment involving expenditures will show that those facilities are systematically less efficient. In this regard, hospitals can be clustered according to some similarities in the inputs and outputs into homogeneous sets. For example, unsupervised machine learning approaches can be used since they do not make any assumptions about which hospitals will be placed in which clusters. This situation is investigated in Contribution 2. Results of this contribution indicate that hospital heterogeneity may account for some variation in efficiency scores of the German hospital market. The performance of a hospital may be influenced by the economic and legislative conditions of its location. Ignoring these environmental variables will result in biased performance measurements. Inclusion of environmental variables may overcome this problem. In general, these might be related to the level of support the hospital receives from the catchment area. Some organizations, particularly those that provide a service, may even find it difficult to determine their catchment area. However, when it comes to defining and measuring environmental variables, the addition of those variables leads to new pitfalls (see Dyson et al 2001).

In spite of all these issues, DEA continues to gain popularity as a method of assessing health care providers. However, the stochastic frontier approach, simple ratios, fixed-effect models, and other methods are all alternatives to estimate best practices. The reason for this is that DEA offers many advantages when it comes to evaluating the performance of health care organizations. In the first place, they are nonparametric, which means no functional form (e.g., nonlinear, log-linear, etc.) has to be specified explicitly. Second, DEA measures best practices by comparing hospital performance to that of all other hospitals in the sample. By doing so, DEA is able to identify the source and size of performance shortcomings. This differs from statistical regressions averaging many hospitals' performance together. A third advantage is that unlike regression and other statistical methods, DEA allows for multiple variables, so the overall results are presented in a single, consolidated measure. For the last advantage, DEA groups hospitals into comparable subgroups to identify those that achieve the best results. Hospitals at the frontier perform the best and are considered efficient. Inefficient hospitals are those that are not located at the frontier; their efficiency is determined by distance from the frontier.

3.3 *Question 3: In measuring teaching hospital efficiency, what should be considered?*

A teaching hospital is known for its advanced level health care, high concentration of resources, and complex processes. In addition to providing care, they are deeply engaged in teaching and research as

well. Costs for these hospitals are normally higher than those for their non-teaching counterparts. This means that they should be more integrated with the circumjacent health care system. In addition, there is a need to develop new management studies to reduce resource waste. Health care organizations have had difficulty finding a single metric to measure performance. Since these services typically aim to meet multiple, intangible, conflicting, ambiguous, and complex goals. Therefore, any study of hospital efficiency is subject to criticism for not taking into account clinical innovation, quality of services, or evolving service characteristics. To date, researchers have focused most of their attention on acute hospitals. Most of these studies focused solely on DEA application. They mostly use the CCR model for measuring overall technical efficiency. Having proper inputs and outputs that are able to adequately describe teaching hospital activities and services is a critical aspect of complexity. Besides measuring care outputs, teaching hospital efficiency measurement is primarily concerned with capturing academic outputs such as teaching. This academic function of the teaching hospital can be represented by counting the number of students, the number of graduates, and third-party funding income in the efficiency analysis process. The academic function of a teaching hospital can be measured by counting the number of students, the number of graduates, and the income from third parties.

However, these measures present some complexity by their very nature, such as the integrity of the number of graduates. By definition, DEA studies assume that inputs and outputs are real (continuous) values. The reality is that there are many situations in which some inputs or outputs are unavoidably integer values. It is inaccurate to consider the number of beds (as an input) and outpatients (as an output) in hospital efficiency studies as real. PPS is a convex combination of evaluating units in both the CRS and VRS DEA models without essentially considering integrality constraints associated with some inputs and outputs. In the case of large integer values, imposing integrality constraints by rounding off the optimum solution may not make a significant difference in the optimality, but this is not the case for small integer values where a few units less or more can make a major difference. A real-valued assumption about integer inputs and outputs may lead to infeasibility (e.g., operations outside of the PPS) or to a dominated (inferior) operating unit. Another fundamental assumption of traditional DEA models is categorizing measures (factors) as inputs or outputs. There are some situations, however, where some measures can be viewed as either inputs or outputs in teaching hospital efficiency studies. Consider, for instance, third-party funding income. In the efficiency evaluation of university hospitals, this measure may be construed as an input (a type of revenue received) or as an output, since funding agencies tend to allocate funds to hospitals with the greatest impact. In the area of measuring teaching hospital efficiency, both situations can exist simultaneously, i.e., some measures can serve as inputs as well as outputs and take integer values only. Consider, for instance, the number of graduates in a university hospital. Depending on their contribution, they may represent inputs (available human resources for the hospital) or outputs (trained staff, a result of teaching/research funding). Therefore, traditional DEA hospital studies probably would not have been useful to managers

in the field. An SBM DEA model can then be constructed which can account for the integer nature of certain measures whose status can be handled in a flexible manner, such as the model outlined in Contribution 3.

3.4 *Question 4: At the crossroads of internationalization, how can we analyze university efficiency?*

Comprehensive internationalization aims to improve the quality of services by exposing recipients to global themes. The integration of internationalization is a standard across all missions and operations to ensure global outcomes in teaching, research, and student programming. Comprehensive internationalization improves outcomes for stakeholders in higher education, including economic beneficiaries. As a result, internationalization has an important impact on university business models since it enhances both prestige and revenues for universities. Internationalization has been shown to improve university performance, so it is vital to determine if internationalization efficiency correlates with the success of university missions.

The relative efficiency of university business model components provides insights into the overall efficiency of university business models. To accomplish university missions more effectively, university business models and internationalization should be studied by demonstrating how the two concepts are interconnected. Policymakers and university administrators can then use these metrics to evaluate aspects of business models and mission accomplishments. It is important to use appropriate mathematical techniques to measure internationalization's relative impact on university mission accomplishments and university business models. Human and financial resources invested by governments and institutions in pursuing their missions contribute to the relative efficiency of university mission accomplishment. These broad institutional ends can be accomplished with the total expenditures (financial and human) and the total academic staff. Traditionally, university missions include teaching, research, and service to society. Total graduates represent the teaching mission, citations indicate the quality of research publications, and patenting, represented by total patent filings, represents the value that patents create for society.

To determine internationalization inputs, comprehensive internationalization literature is useful. International strategies are operationalized by university administrators, represented by total international office staff. Total funding from the EU and other international organizations showing financial inputs for international research and programmatic efforts in university mission pursuit. In analyzing internationalization mission performance, the number of international professors displaying diversity recruitment efforts and internationalization in teaching can be used. Recruiting international students and incoming exchange students contribute to internationalization domestically. It can be expressed as the total number of international students who are enrolled full-time and the total number

of students participating in exchange programs. One of the other outputs at this level concerns university international partnerships. By partnering with institutions and organizations across the world, a university can expand its international influence. When considering internationalization efficiency analysis, some inputs (and outputs) work proportionally, while others are substitutional such as academic staff salaries which are usually included in budgets. Hence, radial models may mislead us when we want to assess university performance with DEA.

Geographical differences could affect both internationalization and educational mission performances, which could, in turn, contribute to university heterogeneity. For instance, universities in affluent regions may benefit from environmental spillover effects. In order to build a strong higher education application, second stage models that examine factors associated with DEA scores should include such non-discretionary factors. In two-stage analysis, the DEA is solved using traditional inputs and outputs, and then the efficiency scores are regressed against the non-discretionary (environmental) variables from the first stage. The estimated regression coefficients can be used to adjust the efficiency scores so that all efficiency scores reflect the same level of environment if there is a significant relationship between environmental factors and DEA estimates. However, this approach is problematic because the efficiency scores are serially correlated with each other. Consequently, it violates the assumption of independent and equal distribution of variables in classical regression. It is not recommended to draw definite conclusions from this analysis by using conventional statistical tests. Instead, it may be viewed as exploratory, indicating which non-discretionary variables seem to impact performance the most.

4 Conclusions

Both the higher education and the health care industries are characterized by similar missions, organizational structures, and resource requirements. There has been increasing pressure on universities and health care delivery systems around the world to improve their performance during the past decade. That is, to bring costs under control while ensuring high-quality services and better public accessibility. Achieving superior performance in higher education and health care is a challenging and intractable issue. Although many statistical methods have been used, DEA is increasingly used by researchers to find best practices and evaluate inefficiencies in productivity. By comparing DMU behavior to actual behavior, DEA produces best practices frontier rather than central tendencies, that is, the best attainable results in practice. The dissertation primarily focuses on the advancement of DEA models primarily for use in hospitals and universities. In Section 1 of this dissertation, the significance of hospital and university efficiency measurement, as well as the fundamentals of DEA models, are thoroughly described. The main research questions that drive this dissertation are then outlined after a brief review of the considerations that must be taken into account when employing DEA. Section 2 consists of a summary of the four contributions. Each contribution is presented in its entirety in the appendices. According to these contributions, Section 3 answers and critically discusses the research questions posed.

Using the Translog production function, a sophisticated data generation process is developed in the first contribution based on a Monte Carlo simulation. Thus, we can generate a wide range of diverse scenarios that behave under VRS. Using the artificially generated DMUs, different DEA models are used to calculate the DEA efficiency scores. The quality of efficiency estimates derived from DEA models is measured based on five performance indicators, which are then aggregated into two benchmark-value and benchmark-rank indicators. Several hypothesis tests are also conducted to analyze the distributions of the efficiency scores of each scenario. In this way, it is possible to make a general statement regarding the parameters that negatively or positively affect the quality of DEA estimations. In comparison with the most commonly used BCC model, AR and SBM DEA models perform much better under VRS. All DEA applications will be affected by this finding. In fact, the relevance of these results for university and health care DEA applications is evident in the answers to research questions 2 and 4, where the importance of using sophisticated models is stressed.

To be able to handle violations of the assumptions in DEA, we need some complementary approaches when units operate in different environments. By combining complementary modeling techniques, Contribution 2 aims to develop and evaluate a framework for analyzing hospital performance. Machin learning techniques are developed to perform cluster analysis, heterogeneity, and best practice analyses. A large dataset consisting of more than 1,100 hospitals in Germany illustrates the applicability of the integrated framework. In addition to predicting the best performance, the

framework can be used to determine whether differences in relative efficiency scores are due to heterogeneity in inputs and outputs. In this contribution, an approach to enhancing the reliability of DEA performance analyses of hospital markets is presented as part of the answer to research question 2. In real-world situations, integer-valued amounts and flexible measures pose two principal challenges. The traditional DEA models do not address either challenge. Contribution 3 proposes an extended SBM DEA model that accommodates such data irregularities and complexity. Further, an alternative DEA model is presented that calculates efficiency by directly addressing slacks. The proposed models are further applied to 28 universities hospitals in Germany. The majority of inefficiencies can be attributed to "third-party funding income" received by university hospitals from research-granting agencies. In light of the fact that most research-granting organizations prefer to support university hospitals with the greatest impact, it seems reasonable to conclude that targeting research missions may enhance the efficiency of German university hospitals. This finding contributes to answering research question 3.

University missions are heavily influenced by internationalization, but the efficacy of this strategy and its relationship to overall university efficiency are largely unknown. Contribution 4 fills this gap by implementing a three-stage mathematical method to explore university internationalization and university business models. The approach is based on SBM DEA methods and regression/correlation analyses and is designed to determine the relative internationalization and relative efficiency of German universities and analyze the influence of environmental factors on them. The key question 4 posed can now be answered. It has been found that German universities are relatively efficient at both levels of analysis, but there is no direct correlation between them. In addition, the results show that certain locational factors do not significantly affect the university's efficiency.

For policymakers, it is important to point out that efficiency modeling methodology is highly contested and in its infancy. DEA efficiency results are affected by many technical judgments for which there is little guidance on best practices. In many cases, these judgments have more to do with political than technical aspects (such as output choices). This suggests a need for a discussion between analysts and policymakers. In a nutshell, there is no doubt that DEA models can contribute to any health care or university mission. Despite the limitations we have discussed previously to ensure that they are used appropriately, these methods still offer powerful insights into organizational performance. Even though these techniques are widely popular, they are seldom used in real clinical (rather than academic) settings. The only purpose of analytical tools such as DEA is to inform rather than determine regulatory judgments. They, therefore, have to be an essential part of any competent regulator's analytical arsenal.

5 References

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.

Charnes, A., & Cooper, W. W. (1962). Programming with linear fractional functionals. *Naval Research Logistics Quarterly, 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.

Charnes, A., Cooper, W. W., & Rhodes, E. (1981). Evaluating Program and Managerial Efficiency: An Application of Data Envelopment Analysis to Program Follow Through. *Management Science, 27*(6): 668–697.

Coelli, T. J., Prasada Rao, D. S., O'Donnell, C. J., & Battese, G. E. (Eds.) (2005). An Introduction to Efficiency and Productivity Analysis. Boston, MA: Springer US.

Cooper, W. W., Seiford, L. M., & Tone, K. (2007). Data envelopment analysis: a comprehensive text with models, applications, references and DEA-solver software. 2nd ed. Boston, MA: Springer US.

Dyson, R. G., Allen, R., Camanho, A. S., Podinovski, V. V., Sarrico, C. S., & Shale, E. A. (2001). Pitfalls and protocols in DEA. *European Journal of Operational Research, 132*(2): 245–259.

Emrouznejad, A., & Yang, G. (2018). A survey and analysis of the first 40 years of scholarly literature in DEA: 1978–2016. *Socio-Economic Planning Sciences, 61*: 4–8.

Farrell, M. J. (1957). The Measurement of Productive Efficiency. *Journal of the Royal Statistical Society: Series A (General), 120*(3): 253–281.

Federal Statistical Office (2021). Statistisches Bundesamt. Federal Statistical Office. Wiesbaden. Available online at https://www.destatis.de/, checked on 10/22/2021.

Kempkes, G., & Pohl, C. (2010). The efficiency of German universities–some evidence from nonparametric and parametric methods. *Applied Economics, 42*(16): 2063–2079.

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.

Liu, J. S., Lu, L. Y.Y., Lu, W.-M., & Lin, B. J.Y. (2013). A survey of DEA applications. *Omega, 41*(5): 893–902.

McAdam, M., Miller, K., & McAdam, R. (2017). University business models in disequilibrium – engaging industry and end users within university technology transfer processes. *R&D Management, 47*(3): 458–472.

OECD, & Union, E. (2020). Health at a Glance: Europe 2020. Available online at https://www.oecd-ilibrary.org/content/publication/82129230-en.

Ozcan, Y. A. (2014). Health Care Benchmarking and Performance Evaluation. Boston, MA: Springer US (210).

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.

Street, A., Smith, P. C., & Jacobs, R. (Eds.) (2006). Measuring Efficiency in Health Care: Analytic Techniques and Health Policy. Cambridge: Cambridge University Press.

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.

Tone, K. (2017). Advances in DEA Theory and Applications: *With Extensions to Forecasting Models*. 1st. Hoboken NJ: John Wiley & Sons Inc (Wiley Series in Operations Research and Management Science).

Valero, A., & van Reenen, J. (2019). The economic impact of universities: Evidence from across the globe. *Economics of Education Review, 68*: 53–67.

Zhu, J., & Cook, W. D. (2007). Modeling data irregularities and structural complexities in data envelopment analysis. New York, NY: Springer.

Appendix I. Analyzing the Accuracy of Variable Returns to Scale Data Envelopment Analysis Models

Mansour Zarrin and Jens O. Brunner

Chair of Health Care Operations / Health Information Management, Faculty of Business and Economics, University of Augsburg, Universitätsstraße 16, 86159 Augsburg, Germany

Status: Submitted to European Journal of Operational Research (Under Review), Category A.

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 develop a Monte Carlo simulation-based data generation process. The process generates an artificial dataset using the Translog production function (instead of 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 most used 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.

Keywords. Monte Carlo Data Generation; Data Envelopment Analysis; Assurance Region; Slacks Based Measurement; Variable Returns to Scale

---

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. 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 et al. (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 et al. 2018). In this framework, BCC (Banker, Charnes, and Cooper) DEA model, introduced by Banker et al. (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 et al. 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.

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) (Resti 2000; Holland

and Lee 2002; Simar and Wilson 2002; Ruggiero 2005; van Biesebroeck 2007; López et al. 2016). 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 Translog[1] 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 and Sicilia 2017). However, models 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 and 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 1,000. 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 (Kohl et al. 2019; Mahmoudi et al. 2020; Kaffash et al. 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 (Simar and Wilson 2015; Banker et al. 2019). 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 (Bogetoft and Otto 2011b; Banker and Natarajan 2011). 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

[1] Translog stands for transcendental logarithmic.

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 estimation[2] distributions for two groups of DMUs through developing statistical tests including parametric and non-parametric ones. For example, Cummins et al. (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 et al. (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 et al. (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, 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

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

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 et al. (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 7,776 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 must 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. Figure 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 performance indicators. Constructing statistical tests allows us to evaluate the null hypothesis of no difference in the distributions of true efficiency ($\boldsymbol{\theta}$) and estimated efficiency ($\widehat{\boldsymbol{\theta}}$) obtained from 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 his 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 $\boldsymbol{\theta}$ and $\widehat{\boldsymbol{\theta}}$ are distributed as normal with parameters ρ_1 and ρ_2, respectively. Then, the test statistic can be calculated as $\left(\sum_j(\theta_j)^2/n\right)/\left(\sum_j(\hat{\theta}_j)^2/n\right)$ 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 and Natarajan 2011). Their 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 $(x_i, i \in \mathbb{M} = \{1, \dots, m\})$ and true efficiency values (θ_j) for each DMU in which the regularity conditions are met. 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 and Otto 2011a).

$$\ln y = \sum_{i=1}^{m} \alpha_i \ln x_i + \frac{1}{2}\sum_{i=1}^{m}\sum_{h=1}^{m} \beta_{ih} \ln x_i \ln x_h \qquad (1)$$

where, 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 Figure 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 $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$, true efficiency ($\boldsymbol{\theta}$), input vector $\boldsymbol{x}$ (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.

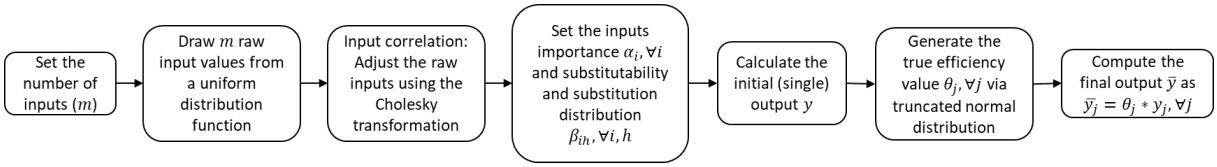


Figure 1. Developed DGP for each artificial DUM

The value of true efficiency (θ) is drawn from a truncated normal distribution and then multiplied with 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 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: $\bar{y}_j = \theta_j * y_j$.

Adjusting the number of inputs, the range of inputs, and the correlation among inputs all lead to the generation of the input vector $\boldsymbol{x}$. 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 region[3] 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 DRS setting (Coelli et al. 1998). We represent units that have exactly the optimal scale of operations as $\boldsymbol{x}^{CRS}$. 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).

$$\Phi_{Oj}(\boldsymbol{x}_j, y_j) = \sum_{i\in\mathbb{M}} \frac{\partial \ln y}{\partial \ln x_i} = \sum_{i\in\mathbb{M}} \alpha_i + \sum_{i\in\mathbb{M}} \left(\beta_{ii} + \sum_{h\in\mathbb{M}\setminus\{i\}} \beta_{ih} \right) \ln x_i \begin{cases} \overset{!}{>} 1 \leftrightarrow IRS \\ \overset{!}{=} 1 \leftrightarrow CRS \\ \overset{!}{<} 1 \leftrightarrow DRS \end{cases}, \forall j \in \mathbb{N} \qquad (2)$$

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

where $\Phi_{Oj}(\boldsymbol{x}_i, y_j)$ represents the scale elasticity value of DMU_j at point $(\boldsymbol{x}_j, y_j)$ as the output distance function. If the value of this function is greater than, equal to, and lower than 1, we respectively have IRS, CRS, and DRS.[4] According to Eq. (2), we can define the sufficient conditions for satisfying the global VRS that still allows the implementation of substitution effects as: $\sum_{i\in\mathbb{M}} \alpha_i > 1 \cap \sum_{i\in\mathbb{M}}(\beta_{ii} + \sum_{h\in\mathbb{M}\setminus\{i\}} \beta_{ih}) \ln x_i < 0$.

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 should set $\sum_{i\in\mathbb{M}} \alpha_i > 1$, then $\sum_{i\in\mathbb{M}} \alpha_i = 1 + \omega,\ \omega > 0$ in Eq. (2). ω is the parameter that 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 through altering the value of α. We here apply two different adjustments containing equal and equidistant importance. In both settings, we must hold $\sum_{i\in\mathbb{M}} \alpha_i = 1 + \omega,\ \omega > 0$ 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).

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

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

Proof. $\sum_{i=1}^{m} \alpha_i = \sum_{i=1}^{m} \frac{1+\omega}{m} = \sum_{i=1}^{m} \left(\frac{1}{m} + \frac{\omega}{m}\right) = \left(m \cdot \frac{1}{m} + m.\frac{\omega}{m}\right) = 1 + \omega \overset{\omega>0}{\rightarrow} \sum_{i=1}^{m} \alpha_i > 1$. ■

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 the 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 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 results due to this regularity.

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

Proposition 2. Eq. (4) fulfills the condition of $\sum_{i\in\mathbb{M}} \alpha_i > 1$.

Proof. $\sum_{i=1}^{m} \alpha_i = \sum_{i=1}^{m} \frac{(1+\omega)\cdot(i+m)}{1.5m^2+0.5m} = \sum_{i=1}^{m} \frac{(1+\omega)\cdot i}{1.5m^2+0.5m} + \sum_{i=1}^{m} \frac{(1+\omega)\cdot m}{1.5m^2+0.5m} = (1+\omega) \cdot \left[\frac{\frac{1}{2}m\cdot(m+1)}{1.5m^2+0.5m} + \frac{m\cdot m}{1.5m^2+0.5m}\right] = (1+\omega) \cdot \left[\frac{1.5m^2+0.5m}{1.5m^2+0.5m}\right] = 1 + \omega \overset{\omega>0}{\longrightarrow} \sum_{i=1}^{m} \alpha_i > 1$. ■

[4] The corresponding output scale efficiency value $SE_{Oj}(x_j, y_j)$ for DMU_j can be calculated by $\ln SE_{Oj}(x_j, y_j) = -(\Phi_{Oj}(x_j, y_j) - 1)^2 / 2\sum_{i=1}^{m}\sum_{h=1}^{m} \beta_{ih}$ (Balk 2001).

The second term of Eq. (2) i.e., $\sum_{i\in\mathbb{M}}\left(\beta_{ii}+\sum_{h\in\mathbb{M}\setminus\{i\}}\beta_{ih}\right)\ln x_i$, 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 $\beta_{ih}\overset{!}{=}\beta_{hi},\forall i,h$ (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 $\sum_{i\in\mathbb{M}}\alpha_i=1+\omega$, the second term of Eq. (2) must be exactly equal to $-\omega$, in other words, $\sum_{i\in\mathbb{M}}\left(\beta_{ii}+\sum_{h\in\mathbb{M}\setminus\{i\}}\beta_{ih}\right)\ln x_i\overset{!}{=}-\omega$ to achieve CRS at $\boldsymbol{x}^{CRS}$, 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., $x_i^{CRS}=x^{CRS},\forall i$).

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

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, $\beta_{ih}\propto\sigma_{ih}*\nu,\forall i,h$. 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, we must hold $\sigma_{ih}\overset{!}{=}\sigma_{hi},\forall i,h$. 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 $\beta_{ih},\forall i,h$. 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 VRS setting through the substitution distribution (σ_{ih}).

For the equal substitution distribution, we have $\sum_{i\in\mathbb{M}}\left(\beta_{ii}+\sum_{h\in\mathbb{M}\setminus\{i\}}\beta_{ih}\right)=m\cdot(\beta_{ii}+\sum_{h\in\mathbb{M}\setminus\{i\}}\beta_{ih})$ by construction, as a result, we can rewrite Eq. (5) as $\beta_{ii}+\sum_{h\in\mathbb{M}\setminus\{i\}}\beta_{ih}=-\frac{\omega}{m\cdot\ln x_i^{CRS}},\forall i$.

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

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

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

$\beta_{ii} + \sum_{h\in\mathbb{M}\setminus\{i\}} \beta_{ih} = \frac{-\nu\cdot\omega}{m\cdot\ln x_i^{CRS}} + \sum_{h\neq i} \frac{(\nu-1)\cdot\omega}{m\cdot(m-1)\ln x_i^{CRS}} = \frac{-\nu\cdot\omega}{m\cdot\ln x_i^{CRS}} + \frac{(m-1)(\nu-1)\cdot\omega}{m\cdot(m-1)\ln x_i^{CRS}} = -\frac{\omega}{m\cdot\ln x_i^{CRS}} \rightarrow$

$\beta_{ii} + \sum_{h\neq i} \beta_{ih} = -\frac{\omega}{m\cdot\ln x_i^{CRS}},\ \forall i.$ ■

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

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

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).

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

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., $\sigma'_{ii} = -\frac{m\cdot\left(1.5-\frac{i-1}{m-1}\right)-\left(2-2\cdot\frac{i-1}{m-1}\right)}{1.5\cdot m-2}, \forall i$ 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 $\sigma'_{ii} + \sum_{h\in\mathbb{M}\setminus\{i\}}\sigma'_{ih} = 0, \forall i$.[5] Now, by replacing these two expressions in Eq. (5) and operating, we have: $\sum_{i\in\mathbb{M}}\left(\beta_{ii} + \sum_{h\in\mathbb{M}\setminus\{i\}}\beta_{ih}\right) = \sum_{i\in\mathbb{M}}\left(-\frac{\omega\cdot(1-\nu\cdot\sigma'_{ii})}{m\cdot\ln x_i^{CRS}} + \sum_{h\in\mathcal{M}\setminus\{i\}}\frac{\omega\cdot\nu}{m\cdot\ln x_i^{CRS}}\cdot\sigma'_{ih}\right) =$

$\sum_{i\in\mathbb{M}}\left(-\frac{\omega}{m\cdot\ln x_i^{CRS}} + \frac{\omega\cdot\nu\cdot\sigma'_{ii}}{m\cdot\ln x_i^{CRS}} + \frac{\omega\cdot\nu}{m\cdot\ln x_i^{CRS}}\cdot\sum_{h\in\mathcal{M}\setminus\{i\}}\sigma'_{ih}\right) = \sum_{i\in\mathbb{M}}\left(-\frac{\omega}{m\cdot\ln x_i^{CRS}}\cdot\left(1+\nu\cdot\sigma'_{ii}+\nu\cdot\right.\right.$

$\left.\left.\sum_{h\in\mathbb{M}\setminus\{i\}}\sigma'_{ih}\right)\right) = \sum_{i\in\mathbb{M}}\left(-\frac{\omega}{m\cdot\ln x_i^{CRS}}\cdot\left(1+\nu\cdot\left(\sigma'_{ii}+\sum_{h\in\mathbb{M}\setminus\{i\}}\sigma'_{ih}\right)\right)\right) = \sum_{i\in\mathbb{M}}\left(-\frac{\omega}{m\cdot\ln x_i^{CRS}}\cdot(1+\right.$

$\left.0)\right) = \sum_{i\in\mathbb{M}}\left(-\frac{\omega}{m\cdot\ln x_i^{CRS}}\cdot(1+0)\right) = \sum_{i\in\mathbb{M}} -\frac{\omega}{m\cdot\ln x_i^{CRS}} = -\frac{\omega}{\ln x^{CRS}}\cdot$ ■

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 and Santín 2009). The partial derivatives of distance functions must satisfy one condition for monotony: for D_O as a single output, all marginal products (f_i) must be non-negative across all inputs (x_i) as outlined by Eq. (10).

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

[5] A detailed derivation of σ'_{ii} and σ'_{ih} can be found in Kohl and Brunner 2020.

$$f_i = \frac{\partial D_O}{\partial x_i} = \frac{\partial \ln D_O}{\partial \ln x_i}\frac{D_O}{x_i} = s_i\frac{D_O}{x_i} \geq 0 \leftrightarrow s_i \geq 0,\ \forall i \tag{10}$$

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).

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

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

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

where, $f_{ih} = \frac{\partial^2 D_O}{\partial x_i \partial x_h} = \frac{\partial f_i}{\partial x_i \partial x_h} = \frac{\partial\left(s_i\frac{D_O}{x_i}\right)}{\partial x_h} = (\beta_{ih} + s_i s_h)\left(\frac{D_O}{x_i x_h}\right), \forall\{i,h|i \neq h\}$, 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 sing. Defining the $i+1$ principle minor by $F(x_i)$, F is negative definite if $(-1)^i\left|F^i(\boldsymbol{x})\right| > 0$.

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.

$$\min / \max \nu \tag{13a}$$

$$s.t. \quad s_i \geq 0, \forall i \tag{13b}$$

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

$$(-1)^i\left|F^i(\boldsymbol{x})\right| > 0, \forall i \tag{13d}$$

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 the 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:

$$\min / \max \nu \tag{14a}$$

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

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

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

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

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

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

We reformulate the model to transform the nonlinear constraints to 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 $-s_1^2 \frac{D_O^3}{x_1^2 x_2^2}(\beta_{22} + s_2^2 - s_2) - s_2^2 \frac{D_O^3}{x_1^2 x_2^2}(\beta_{11} + s_1^2 - s_1)$, is always positive by construction. Consequently, we can simply show that one sufficient condition to fulfill the Eq. (12g) 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:

$$f_{12} = (\beta_{12} + s_1 s_2)\left(\frac{D_O}{x_1 x_2}\right) \geq 0 \leftrightarrow \beta_{12} \geq 0 \tag{15}$$

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 AR[6] (Pedraja-Chaparro et al. 1997), and VRS SBM (Tone 2001). 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).

Table 1. Defined characteristics for generating scenarios

Characteristic	***Value/Level***
Returns to scale	VRS
True efficiencies (θ)	Low, Medium, and High
# DMUs (n)	50, 150 and 450
# Inputs (m)	2, 5 and 7
Importance of inputs (α_i)	SYM and ASYM
Input substitutability (v)	Low and High
Input substitution distribution (β_{ih})	Equal and Unequal
Input range	U[100; 1,100] 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	7,776

The number of DMUs in the generated scenarios can be modified by simply running the DGP for one DMU n times. 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, 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).

[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.

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. Consider 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., $\boldsymbol{x} = [100;\ 1{,}100]$), the obtained values are presented in Table 3. In Appendix B, 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 with the size of the DMU starts at the beginning of the production function. This effect of $x_i^{CRS}$is shown in Figure 2(a) in which we represent the production function of 1,000 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 Figure 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. Figure 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.

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; 1,100]
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

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 $((-1)^i \lvert F^i(\boldsymbol{x}) \rvert > 0)$	$\lvert F^1 \rvert = -0.7671$ and $\lvert F^2 \rvert = 0.3314$

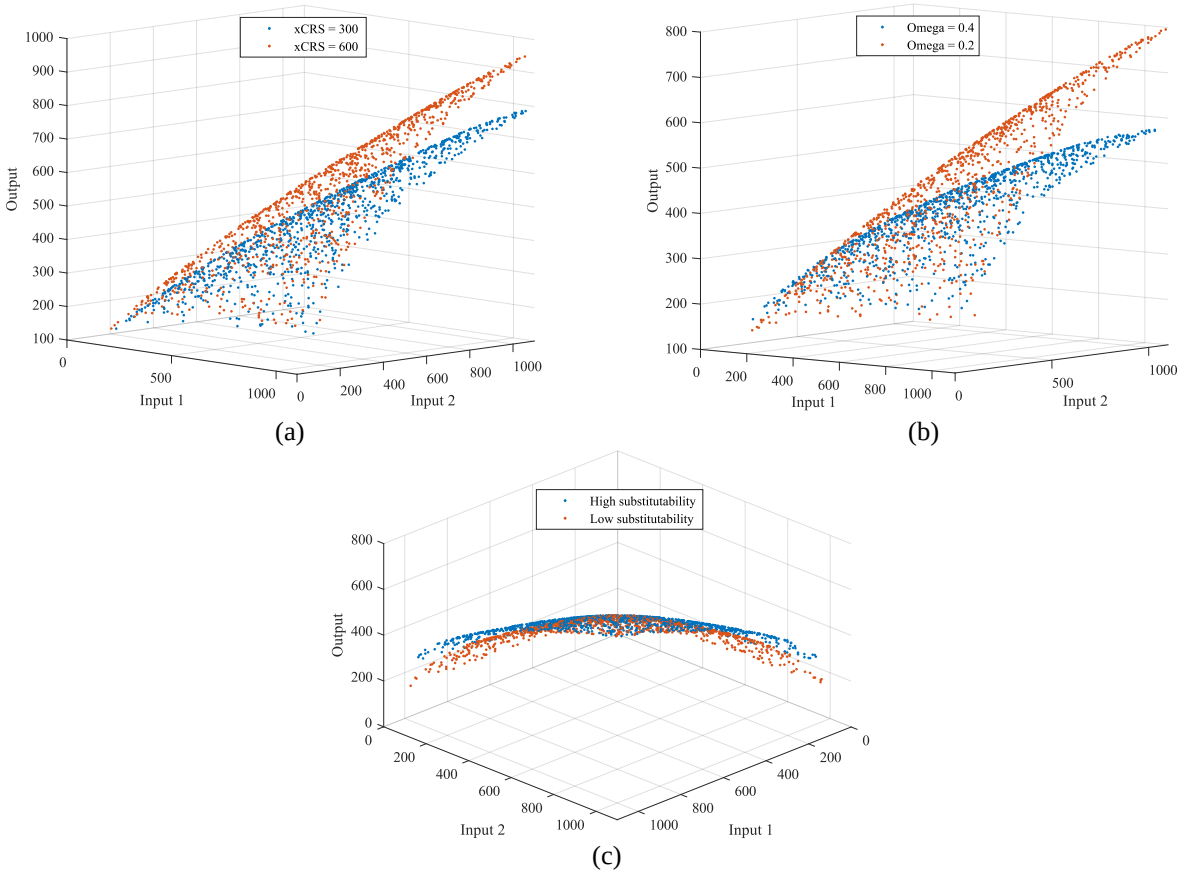


Figure 2. Effect of x_i^{CRS} (a), ω (b), and ν (c) on the form of the production function

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 7,776 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 Figure 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.

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

Indicator	*Statistics*	*Rand*	*CCR DEA*	*BCC DEA*	*AR DEA*	*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

The small value of 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 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 INEFF indicators provide the same result: AR and SBM exhibit high quality and outperform other models.

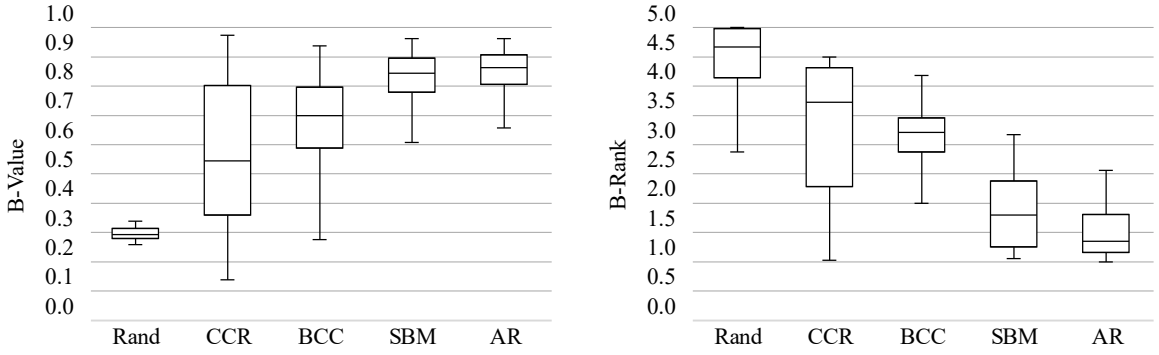


Figure 3. Boxplots of B-Values and B-Ranks obtained from the models

On the basis of Figure 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.

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 C4 and C5 in Appendix C), 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.

Table 5. Results of conducting hypothesis tests

Model	***Number of Rejected Scenarios (%)***
Rand	0 (0%)
CCR	3,930 (50.5%)
BCC	2,368 (30.5%)
AR	870 (11.2%)
SBM	829 (10.7%)

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 C, 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 C3) 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 1,134.

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 C can be used to determine how true efficiency levels affect the quality of 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 DEA models. The quality of the DEA models is marginally diminished 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. The result is that 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 inputs 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 C, 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: $AR \approxeq SBM \succ BCC \succ CCR$. 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.

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 7,776 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 hypothesis tests 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 via 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 in order to ensure acceptable results accuracy.

Consideration of the single-output case is one of the drawbacks of our DGP. The methodology may therefore be generalized to meaningful multi-input multi-output cases in the future. 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 et al. (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.

References

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 William W. Cooper, Lawrence M. Seiford, Joe Zhu (Eds.): Handbook on Data Envelopment Analysis. Boston, MA: Springer US: 273–295.

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 SFA. In Peter Bogetoft, Lars Otto (Eds.): Benchmarking with DEA, SFA, and R. New York, NY: Springer New York: 233–262.

Bogetoft, P., & Otto, L. (2011b). Statistical Analysis in DEA. In Peter Bogetoft, Lars Otto (Eds.): Benchmarking with DEA, SFA, and R. New York, NY: Springer New York: 155–196.

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. Prasada, & Battese, G. E. (1998). Review of Production Economics. In Tim Coelli, D. S. Prasada Rao, George E. Battese (Eds.): An Introduction to Efficiency and Productivity Analysis. Boston, MA: Springer US: 11–37.

Coelli, T. J., Prasada Rao, D. S., O'Donnell, C. J., & Battese, G. E. (Eds.) (2005). An Introduction to Efficiency and Productivity Analysis. Boston, MA: Springer US.

Cordero, J. Manuel, 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. David, 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. In : The Measurement of Productive Efficiency and Productivity Change. New York: Oxford University Press. Available online at https://oxford.universitypressscholarship.com/10.1093/acprof:oso/9780195183528.001.0001/acprof-9780195183528-chapter-2.

Hazewinkel, M. (1992). Encyclopaedia of Mathematics. Dordrecht: Springer Netherlands.

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. Reza (2020). The origins, development and future directions of data envelopment analysis approach in transportation systems. *Socio-Economic Planning Sciences, 69*: p. 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.

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. Available online at https://methods.sagepub.com/book/central-tendency-and-variability.

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 jth DMU ($j \in \{1, \dots, n\}$), respectively. There are two important points to consider when defining the 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 INEFF 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 are able to 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 Mean Absolute Error (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 and Brunner 2020)

Indicator	***Symbol***	***Formula***
Mean absolute error	MAE	$\frac{1}{n}\sum_{j=1}^{n}\lvert\theta_j-\hat{\theta}_j\rvert$
Spearman Correlation Coefficient	SPEAR	$\frac{\sum_j\left(\mathrm{Rg}(\theta_j)-\overline{\mathrm{Rg}(\theta)}\right)\left(\mathrm{Rg}(\hat{\theta}_j)-\overline{\mathrm{Rg}(\hat{\theta})}\right)}{\sqrt{\sum_j\left(\mathrm{Rg}(\theta_j)-\overline{\mathrm{Rg}(\theta)}\right)^2}\sqrt{\sum_j\left(\mathrm{Rg}(\hat{\theta}_j)-\overline{\mathrm{Rg}(\hat{\theta})}\right)^2}}$
Best-performing DMUs	TOP	$\frac{\lvert\{j:\theta_j\geq Q(\varepsilon)\cap\hat{\theta}_j\geq Q(\varepsilon)\}\rvert}{\lvert\{j:\theta_j\geq Q(\varepsilon)\}\rvert}\cdot\left(1-\frac{\max\{\lvert\{j:\hat{\theta}_j\geq Q(\varepsilon)\}\rvert-\lvert\{j:\theta_j\geq Q(\varepsilon)\}\rvert,0\}}{n}\right)$
Worst-performing DMUs	INEFF	$\frac{\lvert\{j:\theta_j\leq Q(1-\varepsilon)\cap\hat{\theta}_j\leq Q(1-\varepsilon)\}\rvert}{\lvert\{j:\theta_j\leq Q(1-\varepsilon)\}\rvert}\cdot\left(1-\frac{\max\{\lvert\{j:\hat{\theta}_j\leq Q(1-\varepsilon)\}\rvert-\lvert\{j:\theta_j\leq Q(1-\varepsilon)\}\rvert,0\}}{n}\right)$
Mean value over the results of the corridor	CORRI	$\sum_{k=1}^{\gamma}\frac{1}{\gamma}\frac{\lvert\{j:\lvert\theta_j-\hat{\theta}_j\rvert\leq k\cdot\delta\}\rvert}{n}$
Benchmark value	B-Value	$\frac{(1-\mathrm{MAE})+\mathrm{SPEAR}+\mathrm{EFF}+\mathrm{INEFF}+\mathrm{CORRI}}{5}$
Benchmark rank	B-Rank	$\frac{\mathrm{rank}(1-\mathrm{MAE})+\mathrm{rank}(\mathrm{SPEAR})+\mathrm{rank}(\mathrm{EFF})+\mathrm{rank}(\mathrm{INEFF})+\mathrm{rank}(\mathrm{CORRI})}{5}$

Appendix B. One Sample Scenario

Table B. 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	1,147.38	0.7879	26	234.00	610.00	474.28	0.7814
2	295.00	964.00	641.16	0.7908	27	259.00	1,015.00	629.63	0.8469
3	122.00	997.00	215.45	0.5526	28	985.00	974.00	1,254.82	0.7422
4	863.00	173.00	264.04	0.5058	29	894.00	583.00	1,172.46	0.9484
5	307.00	948.00	821.23	0.9880	30	816.00	978.00	621.86	0.4032
6	1,092.00	122.00	372.20	0.9502	31	989.00	814.00	1,197.65	0.7737
7	143.00	565.00	323.09	0.7819	32	961.00	362.00	528.62	0.5639
8	1,045.00	310.00	495.89	0.5783	33	628.00	1,002.00	695.69	0.5156
9	1,075.00	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	1,051.00	939.00	1,708.48	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	1,107.77	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	1,383.95	0.8863
15	227.00	311.00	295.18	0.6347	40	997.00	149.00	268.93	0.5748
16	146.00	1,058.00	365.69	0.7916	41	923.00	995.00	1,384.13	0.8365
17	906.00	534.00	1,095.83	0.9283	42	258.00	380.00	520.20	0.9540
18	813.00	715.00	1,023.81	0.7721	43	765.00	835.00	1,250.16	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	1,316.94	0.7868
21	1,091.00	826.00	1,219.96	0.7496	46	773.00	1,079.00	1,494.10	0.9561
22	243.00	275.00	346.09	0.7584	47	356.00	991.00	468.93	0.5017
23	1,093.00	674.00	1,141.91	0.7857	48	660.00	297.00	315.07	0.4310
24	651.00	992.00	1,112.20	0.8106	49	111.00	613.00	198.47	0.5828
25	970.00	1,025.00	1,364.69	0.7938	50	1,090.00	1,083.00	1,573.86	0.8430

Appendix C. Detailed Results of Analysis of Characteristics

Table C1. 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	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	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	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	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	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	0.339	0.259	0.295	0.017	0	0	0
		600	0.335	0.262	0.295	0.017	0	0	0
	Input range	[100; 1,100]	0.339	0.259	0.295	0.017	0	0	0
		[100; 10,100]	0.335	0.261	0.295	0.017	0	0	0
	Extent of scale effects	0.2	0.339	0.262	0.295	0.017	0	0	0
		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
	Input correlation	0	0.335	0.259	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 C2. 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	1,254
		Medium	0.967	0.169	0.576	0.236	0.505	0.500	1,310
		High	0.958	0.140	0.532	0.246	0.527	0.499	1,366
	#DMU	50	0.973	0.154	0.573	0.227	0.577	0.494	1,495
		150	0.974	0.142	0.573	0.240	0.522	0.500	1,352
		450	0.973	0.140	0.575	0.246	0.418	0.493	1,083
	#Inputs	2	0.965	0.164	0.627	0.232	0.402	0.490	1,042
		5	0.934	0.140	0.571	0.238	0.406	0.491	1,052
		7	0.974	0.236	0.523	0.233	0.708	0.455	1,836
	Input Importance	ASYM	0.974	0.140	0.573	0.238	0.508	0.500	1,977
		SYM	0.973	0.148	0.574	0.238	0.502	0.500	1,953
	Input substitution distribution	Equal	0.974	0.140	0.456	0.200	0.717	0.450	2,789
		Unequal	0.973	0.240	0.691	0.214	0.293	0.455	1,141
	Input substitutability	High	0.974	0.146	0.629	0.234	0.409	0.492	1,590
		Low	0.965	0.140	0.518	0.229	0.602	0.490	2,340
	Efficient size	300	0.973	0.140	0.553	0.236	0.540	0.498	2,101
		600	0.974	0.153	0.594	0.238	0.470	0.499	1,829
	Input range	[100; 1,100]	0.974	0.215	0.677	0.231	0.344	0.475	1,337
		[100; 10,100]	0.892	0.140	0.471	0.197	0.667	0.471	2,593
	Extent of scale effects	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	1,360
		0.8	0.950	0.140	0.480	0.229	0.651	0.477	1,687
	Input correlation	0	0.973	0.148	0.576	0.232	0.461	0.499	1,196
		0.4	0.974	0.142	0.574	0.238	0.513	0.500	1,329
		0.8	0.973	0.140	0.572	0.243	0.542	0.498	1,405

Table C3. 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	0.889	0.148	0.634	0.184	0.391	0.488	1,013
		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	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	1,134
	Input Importance	ASYM	0.938	0.141	0.647	0.203	0.310	0.463	1,206
		SYM	0.935	0.138	0.652	0.199	0.299	0.458	1,162
	Input substitution distribution	Equal	0.938	0.138	0.556	0.232	0.462	0.499	1,798
		Unequal	0.935	0.561	0.742	0.096	0.147	0.354	570
	Input substitutability	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	1,606
	Efficient size	300	0.938	0.138	0.634	0.209	0.328	0.469	1,274
		600	0.935	0.146	0.665	0.191	0.281	0.450	1,094
	Input range	[100; 1,100]	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	1,550
	Extent of scale effects	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	1,162
	Input correlation	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 C4. AR DEA Model

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
	Input range	[100; 1,100]	0.963	0.284	0.846	0.123	0.055	0.229	215
		[100; 10,100]	0.954	0.150	0.765	0.214	0.168	0.374	655
	Extent of scale effects	0.2	0.962	0.681	0.869	0.056	0.000	0.000	0
		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
	Input correlation	0	0.961	0.160	0.789	0.176	0.102	0.303	265
		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 C5. 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
	Input range	[100; 1,100]	0.963	0.283	0.828	0.126	0.049	0.216	191
		[100; 10,100]	0.954	0.150	0.756	0.213	0.164	0.370	638
	Extent of scale effects	0.2	0.962	0.645	0.851	0.066	0.000	0.000	0
		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
	Input correlation	0	0.962	0.161	0.776	0.177	0.094	0.292	244
		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

Appendix II. Homogeneity and Best Practice Analyses in Hospital Performance Management: An Analytical Framework

Mansour Zarrin, Jan Schoenfelder and Jens O. Brunner

Chair of Health Care Operations / Health Information Management, Faculty of Business and Economics, University of Augsburg, Universitätsstraße 16, 86159 Augsburg, Germany

The printed version is a pre-print of an article published in *Health Care Management Science*. The final authenticated version is available online at: https://doi.org/10.1007/s10729-022-09590-8.

Status: Published in Health Care Management Science, Category A.

Zarrin, M., Schoenfelder, J. & Brunner, J.O. (2022). Homogeneity and Best Practice Analyses in Hospital Performance Management: An Analytical Framework. Health Care Management Science. https://doi.org/10.1007/s10729-022-09590-8

Abstract: Performance modeling of hospitals using data envelopment analysis (DEA) has received steadily increasing attention in the literature. As part of the traditional DEA framework, hospitals are generally assumed to be functionally similar and therefore homogenous. Accordingly, any identified inefficiency is supposedly due to the inefficient use of inputs to produce outputs. However, the disparities in DEA efficiency scores may be a result of the inherent heterogeneity of hospitals. Additionally, traditional DEA models lack predictive capabilities despite having been frequently used as a benchmarking tool in the literature. To address these concerns, this study proposes a framework for analyzing hospital performance by combining two complementary modeling approaches. Specifically, we employ a self-organizing map artificial neural network (SOM-ANN) to conduct a cluster analysis and a multilayer perceptron ANN (MLP-ANN) to perform a heterogeneity analysis and a best practice analysis. The applicability of the integrated framework is empirically shown by an implementation to a large dataset containing more than 1,100 hospitals in Germany. The framework enables a decision-maker not only to predict the best performance but also to explore whether the differences in relative efficiency scores are ascribable to the heterogeneity of hospitals.

Keywords: Cluster Analysis; Data Envelopment Analysis; Hospital Efficiency Analysis; Artificial Neural Networks; Heterogeneity Analysis

---

1 Introduction

The Federal Statistical Office[1] of Germany reports that the costs of inpatient hospital care amounted to around 91.3 billion euros in 2017, 3.9% higher than in 2016 (87.8 billion euros). Health care costs are driven primarily by hospitals around the world. Because of this, hospitals must constantly monitor and improve their efficiency. Data Envelopment Analysis (DEA) is one of the most effective tools for measuring efficiency, and it is widely used to evaluate the efficiency of decision-making units (DMUs). Nowadays, the use of DEA is rapidly expanding and its usage for hospital efficiency measurement is widely accepted (Kohl et al. 2019). In particular, basic DEA models have two major issues including restrictions by some fundamental assumptions such as homogeneity of DMUs in the dataset (Dyson et al. 2001; Brown 2006) as well as lack of predictive capabilities while they are frequently used as a benchmarking tool. In the following, we introduce these two issues and then explain the main aims of our study.

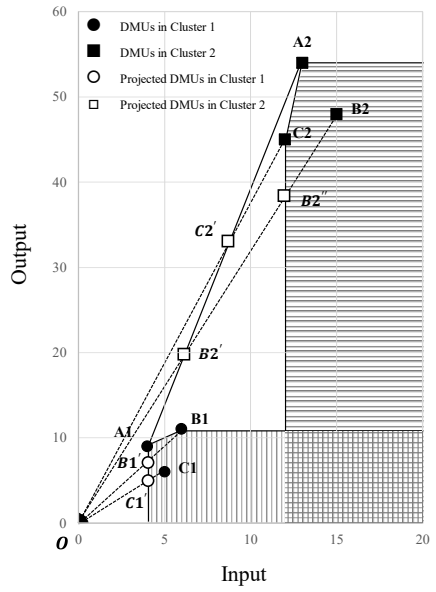
Homogeneity. In the DEA context, homogeneity of a set of DMUs means that all DMUs operate in the same environment and pursue the same target with the same processes. Although significant research has been conducted on the heterogeneity of DMUs, many studies utilize the homogeneity assumption as pointed out by Haas and Murphy (2003) and Wojcik et al. (2019). The applicability of the homogeneity assumption in a sample is usually based on the implicit knowledge of investigators conducting DEA (Dyson et al. 2001). As elucidated by Samoilenko and Osei-Bryson (2010), two factors are important to assume the homogeneity of DMUs in DEA models. The first one that is known as semantic homogeneity brings up the common sense and logic concerned with the meaning assigned to all DMUs in the sample by decision-makers. The second factor is scale homogeneity, where the decision-maker must ensure that the functional similarity of DMUs would not be affected by the input and output levels. Paying no attention to either of these assumptions can heavily influence the results of a DEA application (Dyson et al. 2001). The differences may stem from the type of ownership, the hospital size, and the differences in political and legal environments where the hospitals operate. In the production process, environmental variables are not considered to be traditional inputs and are assumed to be out of the managers' control. The debate about the best ways to incorporate these variables into DEA is still ongoing. Even assuming that the complete consideration of all influential environmental variables is possible, this will cause a lower level of discrimination because of the resulting substantial increase in the number of inputs and outputs (Dyson et al. 2001; Samoilenko and Osei-Bryson 2010).

The impact of the hospital environment can be modeled implicitly by grouping similar DMUs to their transformation capacity (or technology) together. This requires a technique that uncovers categories in the large and multidimensional dataset of DMUs. Incorporating environmental variables in DEA studies has traditionally relied on the two-stage model (Cooper et al. 2011). This approach

[1] Press Release No. 435 as of November 12, 2018

employs the traditional inputs and outputs in the first stage to compute DEA efficiency scores, which are then regressed against the environmental variables (Simar and Wilson 2007). Since both ends of the $0-1$ distribution are restricted, it is often appropriate to use a censored regression model (such as Tobit) for these data. DEA estimates are corrected for environmental effects using regression coefficients. As a result, all efficiency scores will be aligned with the same environment, say the sample mean. However, there is a flaw in this approach. In classical regression, variables are assumed to be independent and identically distributed. According to Simar and Wilson (2007), the DEA efficiency scores considered as the dependent variable in the regression analysis are serially correlated. Therefore, conclusions from the results of this type of study should be drawn with caution. Rather, the method can be regarded as exploratory, indicating which environmental variables are most influential in performance. Another acknowledged approach (Brown 2006; Dyson et al. 2001) to address this issue is to cluster the DMUs into homogenous sets according to some similarities in their environment. Using cluster analysis, we can identify homogeneity between different clusters based on their similarity.

To illustrate how clustering may improve efficiency estimates, consider a sample of 6 DMUs that use an input to generate one output, as shown in Figure 1. DEA benchmarks actual DMU behavior against a set of best practice frontiers. These frontiers create the production possibility set (PPS). As a measure of overall performance, the distance from the DMUs to the frontier is calculated. Best practices, therefore, play a prominent role in calculating the efficiency score. Figure 1 below shows the differences between three different PPSs. As we perform a DEA to measure the efficiency of all six DMUs together, DMUs $A1$ and $A2$ create the efficient frontier. The PPS consists of the area enclosed by this efficient frontier line, plus the horizontal line that extends down from $A1$ and the vertical line that extends right from $A2$. The four DMUs $B1$, $C1$, $B2$, and $C2$ are identified by the DEA as inefficient, and their efficiency can be evaluated by referring to the frontier lines. The efficiency of $B1$, for example, within this PPS is evaluated by $\overline{OB1'}/\overline{OB1} = 0.73$. This unit is inefficient since it underperforms compared to the set of efficient DMUs: $\{A1, A2\}$. It is referred to as the *reference set* or *peer group* of the DMU $B1$. Nevertheless, when we implement clustering before running the DEA, two distinct clusters are detected: cluster 1 (vertical stripes area) includes $A1$, $B1$, and $C1$, and cluster 2 (horizontal stripes area) includes $A2$, $B2$, and $C2$. In cluster 1, the efficient frontier is formed by $A1$ and $B1$, the DMU that was previously shown to be inefficient. $C2$, the DMU that was previously indicated as inefficient, now forms the efficient frontier of cluster 2 together with $A2$. This example illustrates how the clustering can contribute to the estimation of efficiency behind identifying similar DMUs forming the PPS. Clustering may be a useful approach for determining homogeneity and heterogeneity in data sets. To help identify homogenous groups, clustering techniques maximize homogeneity within a group and heterogeneity between groups. Therefore, the resulting inefficiency scores will not be influenced by, e.g., economies of scale.



DMU	$PPS_{\{A1,A2\}}$	$PPS_{\{A1,B1\}}$	$PPS_{\{C2,A2\}}$
A1	*Efficient*	*Efficient*	–
B1	$\frac{\overline{OB1'}}{\overline{OB1}} = 0.73$	*Efficient*	–
C1	$\frac{\overline{OC1'}}{\overline{OC1}} = 0.80$	$\frac{\overline{OC1'}}{\overline{OC1}} = 0.80$	–
A2	*Efficient*	–	*Efficient*
B2	$\frac{\overline{OB2'}}{\overline{OB2}} = 0.78$	–	$\frac{\overline{OB2''}}{\overline{OB2}} = 0.82$
C2	$\frac{\overline{OC2'}}{\overline{OC2}} = 0.93$	–	*Efficient*

Figure 1. Contribution of clustering to measuring efficiency

Traditional DEA models can present several traps for the unwary because of the issue of homogeneity. By analyzing the transformative capacity of hospitals, this study aims to examine the source of differences in the inefficiency of hospitals.

Predictive capabilities. When managers of inefficient hospitals receive the results of a DEA, they usually have subsequent requests, including the possibility of keeping a watchful eye on progress by analyzing what-if scenarios during operational phases and setting target performance levels. Therefore, hospitals must be capable of setting up actionable targets that are specific and measurable. Additionally, analyzing hypothetical scenarios via an adaptive estimation capability can be a valuable addition to assist managers in the monitoring process during the operational phase of change. Although there have been successful models to measure the comparative efficiency of competing units, little attention has been given to including predictability in the performance measurement framework (Kohl et al. 2019). As a second objective, this study explores what level of improvement is needed to see an inefficient hospital become efficient by approximating the efficient frontiers of each cluster and predicting the best performance of each inefficient hospital within its cluster (compared to its leader). Additionally, it facilitates the controlling process during implementation by adding value to if-then scenarios.

2 Literature Review and Contribution

This section reviews the literature relevant to DEA models, neural networks in DEA, clustering in DEA, and the hypothesis tests developed for comparing two groups of DMUs. This section also summarizes our contribution to the literature.

Model Selection. The basic DEA model introduced by Charnes, Cooper, and Rhodes, known as CCR, evaluates the relative efficiency of a set of DMUs (Charnes et al. 1978). Using a variable return-to-scale (VRS) setting, Banker et al. (1984) advance the CCR model. This model is called the BCC model. As radial models, CCR and BCC deal with proportional changes in outputs or inputs. Using these models, the efficiency score is the proportional maximum output (or input) expansion (or reduction) ratio common to all outputs (or inputs) (Tone 2017, 2001). The assumption that these factors will behave proportionally is too restrictive in real-world situations. A further limitation of radial models is ignoring slacks in calculating efficiency scores. Non-radial Slacks-Based Measure (SBM) models have been developed to address these restrictions. SBM DEA models do away with the proportional change assumption and deal directly with slacks. The DEA model has been recognized to be a powerful tool for performance analysis and benchmarking, spanning a wide range of industries and functional areas, including healthcare (Kohl et al. 2019; Almeida Botega et al. 2020; Araújo et al. 2014). In a recent study on the German hospital market, Schneider et al. (2020) investigate hospital urgency scores (noting the average level of medical urgency in all cases treated at a hospital) are compared to technical efficiency. They use the data of 1,428 hospitals throughout Germany for the years 2015, 2016, and 2017. Simar and Wilson (1998) promote bootstrapping as a resampling method for DEA, which has become one of the most commonly used methods in hospital DEA applications (Kohl et al. 2019). There are two main reasons why it is relevant to DEA. DEA estimates tend to be positively biased (Nedelea and Fannin 2013; Mitropoulos et al. 2014) because the estimated production frontier is determined by the units included in the sample. A DMU does not use every input/output combination that is theoretically possible. Hence, the estimated frontier of efficient DMUs is typically too low, even if efficient DMUs are not missing for other reasons (Simar and Wilson 2004). DEA, therefore, assigns efficiency scores that are biased upward because the DMUs are assumed to be closer to the production frontier than they actually are. This upward bias can be corrected via the bootstrapping procedure by creating significance intervals for the efficiency estimates. Our study uses an input-oriented SBM DEA model, in contrast to previous studies (Kwon 2017; Samoilenko and Osei-Bryson 2010, 2008; Omrani et al. 2018), which mostly utilized radial models. We conduct a statistical analysis to determine whether the SBM estimates are significantly biased upward in comparison to the bootstrapped DEA model.

DEA and Machine Learning. Few studies have attempted to reinforce DEA models with machine learning such as artificial neural networks (ANNs) for hospital performance evaluation despite the established effectiveness of these approaches (Kohl et al. 2019). Generally, incorporating ANNs with DEA can be categorized into two distinct research streams. The first consists of studies comparing DEA to ANN as an alternative way of assessing efficiency (Athanassopoulos and Curram 1996; Santín et al. 2004). According to the second stream of research, ANN can be used as a complement to DEA to gain potential advantages. Clustering is one of the machine learning methods used in the literature for subdividing a dataset of DMUs into subsets (clusters) according to how similar the observations are

within each cluster. Several algorithms have been developed in the literature for conducting clustering (Saxena et al. 2017). Among these techniques, three general approaches comprising hierarchical, two-step, and partitional clustering have been used as complements to DEA to handle the scale heterogeneity of samples in the dataset (Mahmoudi et al. 2019; Omrani et al. 2018; Samoilenko and Osei-Bryson 2010). The application of clustering in the literature can be divided into two approaches. One approach is applying clustering to the results of a DEA to facilitate creating multiple reference subdivisions from the original set of DMUs (Bojnec and Latruffe 2008). Second, each DMU is compared with only a subset of its reference set. In the presence of dataset heterogeneity, we can use this approach to isolate the multiple homogenous subsets (Herrera-Restrepo et al. 2016; Samoilenko and Osei-Bryson 2010). In clustering, it is also important to specify the appropriate number of clusters. The quality of partition and cluster validity has been assessed by several authors using different indices (Rocci and Vichi 2008). The Caliński-Harabasz index (CH-index), the Silhouettes, and the Davies-Bouldin criteria were found to be acceptable in a study of clustering conducted by Łukasik et al. (2016). In the literature, details regarding these two criteria and how they are calculated can be found, for example, in Ünlü and Xanthopoulos (2019).

Efficiency Comparison. This study advances the benchmarking paradigm suggested by Samoilenko and Osei-Bryson (2010), which is an extension of Samoilenko and Osei-Bryson (2008), by successfully integrating the clustering and ANN prediction models into an SBM DEA. In Samoilenko and Osei-Bryson (2010), the averages of the relative efficiencies of clusters are used to analyze heterogeneity. A cluster that has a higher average efficiency is referred to as a leader, and a cluster with a lower average efficiency is referred to as a follower. Their method is imprecise because they compare DEA estimates using the mean value of the efficiency scores without considering the distribution of the estimates. The mean value becomes an inappropriate measure when the frequency distribution of the efficiency scores is skewed (Weisberg 1992). Several studies have been conducted where DEA estimation distributions between two groups of DMUs are compared by developing both parametric and non-parametric statistical tests. Banker et al. (2010) develop two sets of parametric and three non-parametric tests. The idea of comparing two groups of DMUs is combined with a heterogeneity analysis in our study. Additionally, we apply our framework to a setting with more than one pair consisting of one leader and one follower.

Our contribution proposes an analytical framework consisting of three stages. We design SOM-ANN for clustering, followed by an SBM DEA model that calculates the relative efficiency of the clustered hospitals. We develop two MLP-ANNs to generate: (i) the transformative capacity model (TCM) to analyze the homogeneity, and (ii) the best practice model (BPM) to predict the level of improvement desired, to achieve efficient operation. The rest of the paper follows this structure. In Section 3, we describe the research methodology and the multi-stage analytical framework combining SOM-ANN, SBM DEA, and MLP-ANN. The dataset of German hospitals used to demonstrate the

framework's applicability is presented in Section 4. The results of the implementation of the framework are presented in Section 5. Section 6 concludes with a discussion of future research directions and conclusions.

3 Methodology

In this section, we describe our proposed framework (see Figure 2). The framework contains three main stages: 1) Clustering using SOM-ANN, 2) efficiency analysis, and 3) heterogeneity and predictability analyses. Each stage is described in detail in the following subsections.

3.1 *Stage* 1: *Cluster Analysis*

We use an SOM-ANN architecture because SOMs are non-linear techniques that can summarize and analyze numerous aspects of variability in a complex, large, multivariate, multi-dimensional dataset (Hudson et al. 2011). In contrast to more traditional clustering methods (such as K-means), SOM-ANN, without imposing a structure on the input/output variables, identifies natural groupings by producing a succinct organization based on similarities among the transformation capacity. As network optimization remains a challenging task, SOM-ANN settings such as initial neighborhood size, topology, and distance functions have been determined by trial and error (Emrouznejad and Shale 2009; Kwon 2017). We also study alternative clustering approaches that are based on the hospitals' natural characteristics: their size (number of beds) and ownership type. The size-related clusters are: small ($beds < 500$), medium ($500 \leq beds < 1{,}000$), and large ($beds > 1{,}000$), while the ownership type clusters are: public, non-profit, and private. This allows us to determine whether natural clustering produces high-quality clusters for hospitals and, consequently, ensures homogeneity within those clusters by comparing the quality indicators calculated for SOM clustering and natural clustering. The function developed for our clustering approach in Python 3.8 is presented in Appendix A.

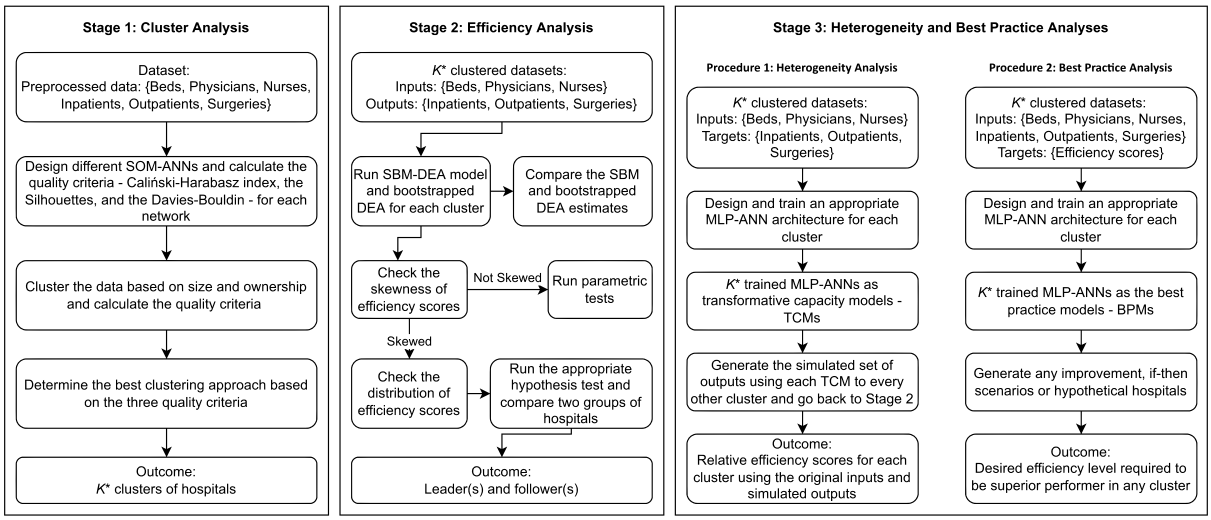


Figure 2. Proposed analytical framework

3.2 *Stage* 2: *Efficiency Analysis*

We run the input-oriented SBM DEA model under VRS settings to calculate the efficiency score of each hospital in each cluster. The mathematical formulation is presented in Appendix B. We also provide details regarding how to calculate the projections based on the slacks determined by the SBM DEA model. In DEA applications, the orientation is chosen based on which parameters managers have more control over (Cooper et al. 2004). While marketers, referral sources, and other methods such as reputation management, can sometimes generate additional patients for hospitals (Ozcan 2014), we use an input orientation under the assumption that hospital managers can more readily control the resources used for patient treatments. Thus, we are interested in the amount by which the resources/inputs (e.g., staff) can be reduced proportionately without reducing the number of treated patients. The downside of using an input-oriented model is the limited applicability when demand for care is higher than the supplied capacity. While this situation may occur for specific treatment types such as chemotherapy or respiratory assistance temporarily, the German healthcare system is set up to continuously assess long-term capacity requirement projections and to react to demand changes with di-/investments into treatment capacities on a state level, so that supply and demand are balanced in the long run.

Furthermore, to determine whether SBM DEA estimates are biased upward or not, we perform a statistical test analysis (explained in the following subsection) between the SBM DEA estimates and bootstrapped DEA estimates produced by implementing the algorithm developed in Daraio and Simar (2007) with the conduct of 200 bootstrap iterations. For brevity, we will not repeat the steps of the algorithm here, however, the reader may refer to Daraio and Simar (2007) for more details.

3.2.1 *Efficiency Comparison of Two Groups of Hospitals*

A DEA estimator of the production frontier is a fully-fledged statistical methodology (Banker 1993) by which we can construct a variety of statistical tests based on efficiency scores represented as stochastic variables. Appendix C describes the comparison algorithm in detail. After indicating the existence of a statistical difference between G_1 and G_2, we reperform the appropriate tests under the one-tailed null hypothesis to indicate whether the efficiency of G_1 is greater than G_2 or vice versa. Throughout the study, all hypothesis tests are performed with a significance level of 5%. Following this procedure, we label the leader and follower in each pair of hospitals.

3.3 *Stage* 3: *Heterogeneity and Best Practice Analyses*

In this stage, two MLP-ANN architectures are designed in two different ways, which are explained in detail in the following subsections. The first architecture supports the scale heterogeneity analysis and the second one is used to predict the actual output level necessary for an inefficient hospital to be efficient. The MLP-ANN maps complex unknown relationships in the dataset because (*i*) MLP-ANNs have a stochastic learning process, which minimizes the chance of being trapped in local minima, and

(*ii*) there is no necessity to specify and know the relationships within the dataset. This architecture, the multilayer feedforward network, is mostly used with the backpropagation algorithm.

3.3.1 *Heterogeneity Analysis*

A model of transformative capacity for each cluster is generated by creating and training an MLP-ANN. Here, it is proposed that score estimates obtained from DEA can be indirectly employed to investigate the factors influencing relative efficiency scores (Hoff 2007; Samoilenko and Osei-Bryson 2010). The DEA efficiency score calculation, however, is hampered by the unavoidable misspecification of the model when determining which inputs are converted into which outputs. Therefore, the decision-maker needs to know the correct transformation function of inputs into the outputs used for conducting the modeling of these estimated scores by DEA. We generate and analyze the transformative capacity model for cluster k denoted by TCM_k. For each cluster, the designed MLP-ANN is trained using the set of input variables (number of beds, physicians, and nurses) as input nodes and the set of output variables (number of adjusted inpatients, outpatients, and surgeries) as output nodes. This is analogous to the way that input data can be transformed into outputs by a given cluster. Then, we investigate for any leader-follower-pair whether the relative efficiency score of the follower improves when comparing the efficiency score distribution of the follower, using the simulated outputs of the follower employing TCM_k of its leader k. When the efficiency score of the follower improves, there is a reason to recommend that the disparity between the original efficiencies of the leading and following clusters is partly due to the differences in transformative capacity. To analyze the scale heterogeneity (scalability), we use the original inputs and outputs of the follower and the initial inputs and simulated outputs of its leader obtained from the TCM_k (follower k) for any leader-follower pairs. If the efficiency of the leader is still higher than the follower, then we can say that scale heterogeneity plays a part in explaining the disparity between the relative efficiencies of the leading and following cluster. In other words, even with the less efficient process of the transformative capacity (i.e., TCM_k, follower k), the leader remains relatively more effective. Visual description is given in Stage 3 of the framework presented in Figure 1 as "Procedure 1: Heterogeneity Analysis."

3.3.2 *Best Practice Analysis*

The second MLP-ANN architecture is designed to deliver improved estimation precision due to its pattern mapping and learning capabilities as a complementary method to DEA. The objective of this analysis is to investigate the predictive capabilities of ANN when used alongside DEA. To this end, the MLP-ANN architecture is trained based on inputs and outputs of the hospitals in each cluster as the input layer and their SBM DEA efficiency scores (see Stage 2 in Figure 1) as the target nodes. Managers can benefit from this analysis in two different ways. First, in a capital-intensive and competitive environment such as in the hospital setting, the ability to estimate input/output levels beyond the

calculated relative efficiency scores is essential for performance benchmarking in real-world applications (Ozcan 2014). Therefore, the first way this analysis can be used by decision-makers is to estimate the efficiency level that can be reached by using a given level of inputs to produce a given level of outputs. Second, the analysis allows managers to set stepwise improvement goals by utilizing what-if scenarios for each inefficient hospital to become an efficient unit, not only in its cluster but also in other clusters without requiring a new DEA. For example, we conduct further experiments to investigate the potential of the proposed framework based on the leader-follower strategy. While DEA has powerful optimization capabilities and a wide range of applications, it has restrictions when working with new or unobserved data sets. If a new DMU is added to a sample and the DEA model is rerun, the results might be completely different as this new DMU might alter the PPS. Hence, the second way this analysis helps managers is to calculate the relative efficiency score of a new or hypothetical hospital by using BPMs trained to learn efficiency patterns existing in the market. This provides managers with alternative paths leading toward best practices, which typically occur at the planning stage and before implementation. Visual description is also given in Stage 3 of the framework presented in Figure 1 as "Procedure 2: Best Practice Analysis."

4 Data Set and Descriptive Statistics

The proposed framework in this study is examined in the context of a large dataset of hospitals recorded by the Federal Joint Committee2 in Germany in 2017. The raw dataset includes all the hospital quality reports of the reporting year 2017. In this study, the information on standard input and output variables for performance assessment of hospitals (Kohl et al. 2019; Tone 2017) was extracted from these reports. Appendix D provides more details about the data sources and a flowchart of the steps involved in data preprocessing. The processed dataset includes 1,124 hospitals.

Kohl et al. (2019) provide some insights into standard input/output settings in their review of hospital DEA studies. Their report indicates that the parameters most used in hospital DEA applications are beds, nurses, physicians, inpatients, and outpatients. These measures are suitable for describing the service process of a hospital as stated by Ozcan (2014). A hospital's capacity can be measured by the number of beds it has. Physicians and nurses play the main role in the hospital's service process. Therefore, the input factors can be considered as beds (Beds), nurses (Nurses), and physicians (Physicians). In our sample, we use full-time equivalents (FTE) of physicians and nurses. As for the outputs, we use the most common output variables used in the literature (Kohl et al. 2019): the number of adjusted inpatients (Adjusted Inpatients) and the number of outpatients (Outpatients). Patients' conditions need to be considered when evaluating inpatient cases, as not every patient requires the same level of care. Following a prior study on efficiency measuring of the German hospital market (Schneider et al. 2020), we apply the case-mix adjustment based on the relative length of stay for groups of hospital

[2] In German: Gemeinsamer Bundesausschuss. https://www.g-ba.de/

diagnoses (according to the International Classification of Diseases Tenth Revision [ICD-10] codes) as suggested by Herr (2008). The German Federal Statistical Office3 publishes hospital statistics on average lengths of stay for each diagnosis group. In addition to these outputs, we consider the number of surgeries based on OPS-54 codes (Surgeries). This output plays a major role in generating net revenue for hospitals. Table 1 represents some descriptive statistics regarding the inputs and outputs of the hospitals in our dataset.

5 Results and Discussion

This section presents the key experimental results of each stage of the proposed framework. We interpret and explain how far these results support the hypothesis and answer the research questions.

Table 1. Descriptive statistics of inputs and outputs of dataset (after preprocessing)

Statistic	*Beds*	*Physicians*	*Nurses*	*Adjusted Inpatients*	*Outpatients*	*Surgeries*
Mean	386.1	131.7	295.2	20,051.6	39,713.2	16,991.7
Standard Error	10.2	5.0	9.5	634.5	2,486.7	606.6
Median	283.0	79.7	199.6	12,262.1	20,780.0	9,795.5
StD	340.5	168.2	318.3	21,253.4	83,368.1	20,335.6
Kurtosis	9.8	28.3	20.7	15.6	137.6	11.9
Skewness	2.5	4.3	3.6	3.0	9.7	2.8
Minimum	50.0	6.0	11.0	628.8	11.0	1.0
Maximum	3,011.0	2,066.7	3,695.7	204,827.6	1,568,896.0	178,580.0
Sum	434,023.0	147,983.0	331,815.8	22,497,902.8	44,637,688.0	19,098,719.0
Confidence Level (95.0%)	19.9	9.8	18.6	1,244.9	4,879.0	1,190.1

*Including all types of physicians such as specialist, non-specialist, and external in full-time equivalent (FTE) unit.
**Including all types of nurses such as pediatric, geriatric, auxiliary, and general in the FTE unit.

5.1 *Results of Cluster Analysis*

For the optimal number of clusters, we create a list of 54 distinct two-dimensional hexagonal layer topologies. We then run the SOM-ANN for each topology of this list to generate clustering vectors. For each clustering vector, three quality criteria are calculated: CH-index, Silhouettes, and Davies-Bouldin (see Appendix E). We then calculate the quality indicators for the clusters resulting from the size and ownership. The results are presented in Table 2. When compared to the best SOM clustering, size (small: $beds < 500$, medium: $500 \leq beds < 1{,}000$, and large: $beds > 1{,}000$) and the ownership (public, non-profit, and private.) of hospitals provide low-quality clusters. Interestingly, clustering based on ownership is ineffective when identifying homogeneity within a group of hospitals and heterogeneity across groups, yet this approach is adopted often in DEA hospital applications with multiple stages (Ozcan 2014; Jacobs et al. 2006; Herr 2008). In identifying homogenous groups, size (number of beds) clustering performs better than ownership; however, they are both outperformed by

[3] https://www.destatis.de/
[4] Chapter 5 of OPS (Operationen- und Prozedurenschlüssel) which is the German modification of the International Classification of Procedures in Medicine.

SOM. By using SOM-ANN, we have three clusters and can calculate the efficiency scores of hospitals in each cluster.

Table 2. Results of comparing the clustering approaches

Clustering Approach	*No. of hospitals*			*CH-index**	*Silhouette***	*Davies-Bouldin****
Size	Small: 853	Medium: 201	Large: 70	647.35	0.48	1.08
Ownership	Non-profit: 450	Private: 238	Public: 436	25.77	-0.11	7.59
SOM	Cluster 1: 186	Cluster 2: 249	Cluster 3: 689	874.54	0.57	0.76

* A high score is achieved when clusters are dense and well separated.
** The score ranges from -1 for incorrect clustering to $+1$ for dense and well-separated clustering.
*** A value closer to zero indicates a better partition.

5.2 *Results of Efficiency Analysis*

We calculate the efficiency of each hospital and the projections calculated for each hospital using an input-oriented SBM DEA under the VRS setting. SBM DEA estimates (G_{SBM}) are compared to bootstrapped DEA estimates (G_{BT}) produced by the implementation of the algorithm developed by Daraio and Simar (2007) to determine if they are biased upward. Table 3 presents the results of the comparison. In all three clusters, efficiency scores are skewed. They follow neither an exponential nor a half-normal distribution.

Table 3. Comparison of bootstrapped DEA and SBM estimates

Cluster	***Mean* (Bootstrapped DEA, SBM)**	***StD* (Bootstrapped DEA, SBM)**	***Median* (Bootstrapped DEA, SBM)**	***p*-value (H_0: $G_{SBM} = G_{BT}$; H_1: $G_{SBM} \neq G_{BT}$)**
1	(0.8078, 0.8300)	(0.1066, 0.1364)	(0.8259, 0.8465)	0.5540
2	(0.6439, 0.6862)	(0.1295, 0.1760)	(0.6469, 0.6575)	0.5650
3	(0.6797, 0.6891)	(0.1259, 0.1716)	(0.6808, 0.6610)	0.5332

Mann–Whitney tests reveal that the distribution underlying input-oriented SBM estimates is not significantly different from the distribution underlying bootstrapped DEA estimates. The *p*-values indicate that the null hypothesis should be retained. We then continue our analysis using the input-oriented SBM DEA model. Table 4 summarizes the results of the relative efficiency scores calculated for the clusters and all hospitals. As a result of clustering, both the mean and median efficiency scores as well as the number of efficient hospitals increase. Table 5 shows that the amounts by which inputs need to be reduced proportionately (while keeping the outputs constant) are significantly diminished after applying cluster analysis. For example, the number of beds that hospitals need to reduce, on average, to become efficient before clustering is 60% higher than after clustering. Clustering all hospitals in one group may conceivably distort the results since an important assumption of DEA is that all DMUs are homogenous.

Table 4. Descriptive statistics of efficiency scores before and after clustering

Statistics	***Cluster 1***		***Cluster 2***		***Cluster 3***	
	Before clustering	*After clustering*	*Before clustering*	*After clustering*	*Before clustering*	*After clustering*
Mean	0.7135	0.8300	0.6034	0.6862	0.5964	0.6891
Standard Error	0.0124	0.0100	0.0108	0.0112	0.0071	0.0065
Median	0.6898	0.8465	0.5905	0.6575	0.5633	0.6610
StD	0.1688	0.1364	0.1706	0.1760	0.1865	0.1716
Kurtosis	-0.1618	0.0765	0.4956	-0.5742	0.0077	-0.4841
Skewness	-0.0005	-0.5851	0.4991	0.3601	0.7116	0.3184
Minimum	0.2202	0.3352	0.2161	0.2973	0.1959	0.2516
Maximum	1.0	1.0	1.0	1.0	1.0	1.0
Efficient DMUs	20	39	9	34	41	84

Table 5. Descriptive statistics of input excesses before and after clustering

Statistics	***Beds***		***Physicians***		***Nurses***	
	Before clustering	*After clustering*	*Before clustering*	*After clustering*	*Before clustering*	*After clustering*
Mean	155.92	96.32	50.94	36.98	115.83	90.68
Standard Error	4.45	3.81	1.73	1.65	3.21	3.05
Median	120.25	57.30	34.66	20.42	85.34	63.91
Mode	0.00	0.00	0.00	0.00	0.00	0.00
StD	149.02	127.60	57.97	55.37	107.75	102.09
Kurtosis	28.48	48.18	15.45	21.59	12.93	17.23
Skewness	3.50	4.66	3.19	3.92	2.74	3.17
Maximum	2,062.56	1,982.96	568.81	559.61	1,169.14	1,165.43
Sum	175,254.98	108,266.58	57,253.24	41,565.68	130,189.20	101,929.65

5.3 *Results of Heterogeneity and Best Practice Analyses*

This section presents the results of the last stage of the proposed framework. First, the simulated output sets for each cluster are generated based on the TCMs created by MLP-ANN. The first procedure of Stage 3 is focused on determining: (*i*) whether the relative efficiency score of hospitals in a certain cluster improves if we consider the TCM of other clusters, and (*ii*) identifying the differences that are partially due to scale heterogeneity. The second part of the analysis aims at exploiting the non-linear mapping capabilities of MLP-ANN by using the input and output data of each cluster as input nodes (input layer) and assigning their efficiency scores received from DEA-SBM as target nodes (output layer). We develop both MLP-ANNs using an end-to-end open-source platform called TensorFlow in Python 3.8. We set the mean absolute percentage error (MAPE) as the performance measure due to its scale independence, interpretability, and simplicity. For training, validation, and testing, we use a random data division function. The training function updates weight and bias values based on "Adam", a stochastic optimization method developed by Kingma and Ba (2014). More details regarding the parameters of the developed MLP-ANNs are provided in Appendix F.

5.3.1 *Results of Heterogeneity Analysis*

For each cluster, we design an MLP-ANN to create a TCM ($TCM_k, \forall k \in \{1,2,3\}$). Using the TCMs of the other two clusters, we simulate the output values of adjusted inpatients, outpatients, and surgeries for each cluster. For example, in the case of Cluster 1, we import the actual inputs (*Beds*, *Physician*, and *Nurses*) of this cluster to the TCMs generated for Cluster 2 (TCM_2) and Cluster 3 (TCM_3) to

generate two simulated output sets for Cluster 1. The simulated outputs are then substituted for the actual outputs of Cluster 1, and the new relative efficiency scores are calculated. As a result, we have three sets of efficiency scores for Cluster 1 based on three sets of outputs: the original outputs, simulated outputs using TCM's Cluster 2 (TCM_2), and simulated outputs using TCM's Cluster 3 (TCM_3). $C_k^{TCM_{k'}}, \forall k, k' \in \{1,2,3\}\ and\ k \neq k'$ represents the set of relative efficiency scores calculated based on actual inputs of Cluster k and the simulated outputs obtained from $TCM_{k'}$. The MAPE values calculated for each TCM are presented in Table 6. Surgeries show the highest MAPE value among the outputs, likely because its variance is higher than that of other outputs across all three clusters.

Table 6. Best settings of the designed MLP-ANNs for simulating outputs

Transformative capacity model	***Layers***	***Train:Test:Validation Ratio***	***MAPE of the test dataset***		
			Adjusted Inpatients	*Outpatient*	*Surgeries*
TCM_1	[20, 10, 10]	75:20:5	15%	16%	24%
TCM_2	[20, 10, 10]	80:15:5	7%	10%	14%
TCM_3	[20, 10, 10]	80:15:5	6%	6%	11%

We must first define the leader-follower relationship for all cluster pairs by comparing the efficiency of two groups of hospitals. The efficiency scores of all clusters are skewed, as shown in Table 4. Following that, according to the algorithm developed for comparing efficiencies, we check whether the efficiency is distributed exponentially or half-normally for each pair of hospitals (G_1 and G_2). Based on the Q-Q (Quantile-Quantile) plots of all clusters, they do not appear to have come from populations with an exponential or half-normal distribution. Therefore, we conduct the Mann–Whitney test to determine if one hospital cluster is stochastically more efficient than the other, i.e., determining the leader and the follower of the pair. Table 7 shows the results of comparing the distribution of efficiency scores of all clusters, including their leader and/or follower. There is no significant difference in efficiencies underlying Clusters 2 and 3. Therefore, in this pair, no leader (or follower) can be identified. However, if we only compare the mean efficiency scores (see Table 4) and determine the leader solely based on them, Cluster 3 emerges as the leader. In this regard, comparing the efficiency of two groups of hospitals only based on mean values could lead to the wrong detection of leaders. Based on the Q-Q plots of the simulated outputs, the new efficiency score sets are neither exponentially nor half-normally distributed. Therefore, we compare efficiency scores using the Mann–Whitney test (see Table 7).

Transformative capacity. We utilize the actual inputs and the simulated outputs of the follower using TCM of its leader and compare the resulting efficiency scores with the original efficiency of the follower. Consider the results reported in Table 8 for Clusters 1 and 2 as one instance. Cluster 1 is the leader of Cluster 2. The results indicate that the efficiency of Cluster 2 as a follower, based on its actual inputs and the TCM_1 outputs ($C_2^{TCM_1}$), has increased compared with its initial efficiency score, i.e., $C_2 < C_2^{TCM_1}$. This means that the difference between the relative efficiencies of Cluster 1 (leader) and Cluster 2 (follower) is caused by the disparities in their transformative capacities. However, this conclusion is

not valid for Cluster 3 ($C_3 > C_3^{TCM_1}$) as the other follower of Cluster 1. For the pair $\{C_2, C_3\}$, whose leader (follower) cannot be identified, this analysis should not be conducted. If we compared the mean values, the leader-follower analysis would proceed as follows: the average efficiency score $C_2^{TCM_3}$ is equal to 0.8838, a significant increase from the initial average efficiency score (0.6862). Thus, we could infer that the disparity in efficiency scores has to do with their differences in transformative capacity. However, as no leader/follower was identified in the first place, the efficiency distributions of the two clusters could not be determined to be significantly different. We can conclude that there are instances where the difference between the relative efficiencies of hospitals in Germany is due to disparities in their transformative capacities.

Scale heterogeneity (scalability). We compare the original efficiency of a follower with the efficiency scores of its leader (based on the actual inputs and the simulated outputs by the TCM of the follower). Results are reported in Table 8. In the case of Clusters 1 and 2, the distributions of the initial efficiency score of the follower (C_2) and the distributions of the efficiency score calculated based on TCM_2 for the leader ($C_1^{TCM_2}$) are compared. Since $C_1^{TCM_2}$ is greater than C_2, Cluster 1 remains the leader of Cluster 2. Thus, scale heterogeneity partially explains the difference in relative efficiencies between Clusters 1 and 2. For Cluster 3, the other follower of Cluster 1, similar results can be observed ($C_3 < C_1^{TCM_3}$). Overall, there is no case in which the relative efficiency score of the leader is smaller than the relative efficiency score of the follower. There is no case in which the new relative efficiency scores of a leader are stochastically lower than those of the follower. In this way, we can argue that a part of the reason for the disparities between the relative efficiency scores of followers and leaders is scale heterogeneity. This indicates that in the German hospital market, despite the less efficient process of TCM (i.e., follower), the leading hospitals are relatively more efficient than the following ones.

Table 7. Comparing relative efficiency scores via Mann–Whitney test

Pair $\{G_1, G_2\}$	***p-value*** $(H_0: G_1 = G_2, H_1: G_1 \neq G_2)$	***Result of hypothesis tests***	***Leader***
$\{C_1, C_2\}$	0.0000	$C_1 > C_2$	C_1
$\{C_1, C_3\}$	0.0000	$C_1 > C_3$	C_1
$\{C_2, C_3\}$	0.6785	$C_2 = C_3$	–

Table 8. Results of comparing relative efficiency scores calculated based on the TCMs via Mann–Whitney test

Analysis	***Leader***	***Follower***	G_1	G_2	***p-value*** $(H_0: G_1 = G_2; H_1: G_1 \neq G_2)$	***Result of hypothesis tests***
Transformative Capacity	1	2	C_2	$C_2^{TCM_1}$	0.0002	$C_2 < C_2^{TCM_1}$
	1	3	C_3	$C_3^{TCM_1}$	0.0000	$C_3 > C_3^{TCM_1}$
Scale Heterogeneity	1	2	C_2	$C_1^{TCM_2}$	0.0164	$C_2 < C_1^{TCM_2}$
	1	3	C_3	$C_1^{TCM_3}$	0.0000	$C_3 < C_1^{TCM_3}$

5.3.2 *Results of Best Practice Analysis*

Similar to the first procedure, our next step is to find the best settings for the newly designed MLP-ANNs (i.e., BPMs) for our best practice analysis of hospitals. The performance measure of the trained BPMs is reported in Table 9. In each case, a low MAPE indicates a good fit and generalizability.

Table 9. Best settings of the designed MLP-ANNs for best practice analysis

Cluster	***Layers***	***Train:Test:Validation Ratio***	***MAPE of the test dataset***
1	[8, 8]	75:20:5	8%
2	[10, 10]	80:15:5	8%
3	[10, 10]	80:15:5	7%

The frontier function can be viewed as the upper limit of the support of the density of hospitals in the input and output space. On the efficient frontier, concavity and monotonicity assumptions are assumed to be preserved by DMUs. However, the bootstrapped estimates do not necessarily preserve the concave monotone increasing condition. As a result, BPMs are trained based on the SBM DEA estimates where concave monotonic properties of the efficient frontier are preserved (Pendharkar 2005, 2011; Kwon 2017).

To elaborate, we look at one inefficient hospital in Cluster 2, for instance, which has an efficiency score of 0.7422. The SBM DEA projections suggest reducing the number of beds by 27%, physicians by 21%, and nurses by 24%. In terms of output, the projection calls for increasing the number of outpatients by 16%, adjusted inpatients by 5%, and surgeries by 887%, which sounds unrealistic. It is now necessary for the management of this hospital to have a list of possible improvement scenarios that determine what efficiency level can be achieved by using a given level of inputs to provide a given level of outputs. Re-running the DEA for every scenario setting is one option. If, however, we want to keep the PPS unchanged, we cannot consider scenarios with lower reduction rates than those predicted by input projections or higher expansion rates than those set by output projections. By reducing beds by 35% and keeping the remaining factors unchanged, DEA might form a new PPS according to the new data. However, the designed BPM of Cluster 2 (BPM_2) can predict the desired level of this hospital's best performance in any setting without concern over creating a new efficient frontier. Table 10 presents the estimation results on possible improvement scenarios for this hospital and shows the projected efficiency increase that can be achieved by decreasing inputs and/or increasing outputs. As we can see from Scenario 7, the management of the hospital under study does not have to follow the projections derived from the DEA (e.g., unrealistic increasing the number of surgeries by about 900%) to become efficient in the peer group. Compared to SBM projections, these changes sound more realistic and applicable. For varying input levels, the proposed approach can support managers in setting optimal levels of outputs (e.g., the number of adjusted inpatients or outpatients). The same analysis and investigation can be applied to every other inefficient hospital.

Furthermore, we conduct additional experimentation to explore the potential of the proposed framework based on the leader-follower strategy. The results presented in Table 7 can also be utilized to measure hospitals' efficiency within a managerial network. In cases where a leader-follower strategy can be applied, managers of inefficient or weakly-efficient hospitals can utilize the BPM(s) of their leader(s) as well. Consider a hospital that is part of a private hospital group with 15 hospitals in Cluster 2 and 10 hospitals in Cluster 1, which is the leader of Cluster 2. As reported in Table 11, the relative efficiency score obtained from the SBM DEA model for this hospital is 0.5797 based on original inputs and outputs. The projection of this hospital suggests that drastic changes would be required to become an efficient hospital in its Cluster 2: reducing the number of beds by 33%, physicians by 53%, and nurses by 40%, and increasing the number of outpatients, and surgeries by 2% and 35%, respectively. As a result of Scenario 5, we need less reduction in inputs and less expansion of outputs generated by the hospital to become efficient when using BPM_1 (leader).

Table 10. Possible improvement scenarios for an inefficient hospital using its cluster's BPM

Actual inputs and outputs		***Beds***	***Physicians***	***Nurses***	***Adjusted Inpatients***	***Outpatients***	***Surgeries***	***Efficiency***
		256	46.5	172.92	19,474.2	7,175	220	0.7423
Projections		188 (-27%)	36.9 (-21%)	130.97 (-24%)	19,474.2 (0%)	15,085.3 (110%)	2,170.8 (887%)	1.0000
Improvement scenarios	1	-5%	-10%	-5%	0%	10%	20%	0.7462
	2	-10%	-10%	-5%	0%	10%	40%	0.7526
	3	-15%	-15%	-10%	0%	10%	60%	0.7708
	4	-20%	-15%	-10%	5%	20%	80%	0.7964
	5	-25%	-20%	-10%	5%	20%	100%	0.8907
	6	-30%	-20%	-15%	5%	20%	150%	0.9599
	7	-35%	-10%	-15%	10%	30%	150%	0.9958
	8	-40%	-10%	-15%	10%	30%	150%	1.0250
	9	-45%	-10%	-15%	10%	30%	0%	1.0224
	10	-50%	-10%	-15%	10%	30%	0%	1.0374

Table 11. Possible improvement scenarios for another inefficient hospital using its leader's BPM

Actual inputs and outputs		***Beds***	***Physicians***	***Nurses***	***Adjusted Inpatients***	***Outpatients***	***Surgeries***	***Efficiency***
		341.0	130.5	275.2	18,313.5	22,221.0	12,969.0	0.5797
Projections		226.8 (-33%)	61.8 (-53%)	165.2 (-40%)	18,313.5 (0%)	22,717.5 (2%)	17,564.8 (35%)	1.0000
Improvement scenarios	1	-5%	-10%	-5%	0%	0%	5%	0.9055
	2	-10%	-10%	-10%	0%	0%	10%	0.9248
	3	-15%	-15%	-15%	0%	2%	15%	0.9531
	4	-20%	-15%	-20%	0%	2%	20%	0.9717
	5	-25%	-20%	-25%	0%	2%	25%	0.9969
	6	-30%	-20%	-30%	0%	5%	30%	1.0159
	7	-35%	-30%	-35%	5%	10%	35%	1.0472
	8	-40%	-30%	-40%	5%	15%	0%	1.0621
	9	-45%	-30%	-45%	5%	0%	0%	1.0677
	10	-50%	-30%	-50%	10%	0%	0%	1.0891

The results show that a nondiscriminatory standard DEA for all hospitals would fail to account for differences in scale heterogeneity, differences in transformational capacities, and likely other exogenous factors that vary between hospitals of the same group. The non-linear mapping and adaptive prediction capabilities of our trained BPMs allow for the compensation of the lack of predictive

capabilities of standard DEA models, which are still frequently used as benchmarking tools. Therefore, the framework proposed in this study can assist managers in setting any performance targets for their hospitals over time.

6 Conclusions

There are limited economic resources available to hospitals. Therefore, it is essential to determine how the resources are being utilized and whether they are being distributed appropriately. DEA has been used in numerous studies. However, if hospitals operate under different environments, basic DEA alone may not be the best approach and may need some complementary approaches to deal with violations of its assumptions. In this study, we propose a framework for improving the discriminatory and estimation power of DEA. Traditional DEA classifies DMUs in the sample as efficient or inefficient, whereas the proposed framework can account for heterogeneity as a result of the size of the dataset and its ability to transform the data. As complementary to DEA, the framework designs two different architectures of neural networks, namely SOM-ANN and MLP-ANN.

The framework examines the hospital dataset that the Federal Joint Committee of Germany recorded in 2017. To ensure complete accuracy and robustness in calculations, many preprocessing steps are involved in each stage of the framework due to the vast and complex dataset. The proposed framework possesses improved prescriptive capabilities over DEA approaches in a heterogeneous environment. The developed framework may also contribute to the creation of continuous improvement opportunities by promoting the best management practices within a group of hospitals. The proposed framework advances the current benchmarking paradigm of hospitals by learning the optimal performance pattern of hospitals on the efficient frontier of each group. By using what-if and identifying improvement scenarios, the framework can assist decision-makers in evaluating efficiencies. There are clearly defined stages in this study's framework, and different methods are employed as part of each stage. Analyzers can address the effect of environmental variables on heterogeneity without adding additional variables to DEA models. The key findings of this study can be summarized as follows:

- Natural clustering of hospitals (i.e., based on ownership or size) would not reveal homogeneity within groups of hospitals, nor would it identify heterogeneity between groups of hospitals.
- According to the SBM DEA estimates, the distribution underlying the bootstrapped DEA estimates is identical to the distribution underlying the SBM DEA estimates.
- The differences in the relative efficiency of some German hospitals can be due to differences in their transformation capacities rather than inefficient input usage in producing outputs. Furthermore, a part of the reason for the disparities between the relative efficiency scores of hospitals is scale heterogeneity.

- The trained BPMs can compensate for the lack of predictability of standard DEA models due to their nonlinear mapping and adaptive prediction abilities.

Most studies ignore the heterogeneity pitfall even though it is widely recognized that DEA studies can be compromised by it. DEA would be more robust if methods were developed to prove the reliability and correctness of results. DEA models alone cannot resolve the major problems in hospital performance management that arise from operating in an environment heterogeneous in nature. Because exogenous factors are complex and multiplicative, identifying and measuring them is challenging. Consequently, the process of selecting a reference set for every hospital should be handled cautiously. As demonstrated by well-established quality indicators, it is interesting to note that, contrary to previous findings (Tiemann et al. 2012; Herr 2008), clustering hospitals based on ownership failed to create homogeneity within a group and heterogeneity between groups of hospitals under study. The findings are also different from what one would intuitively expect to find in the context of performance management of hospital markets. For example, one could assume that the relative homogeneity of hospitals would allow for simple emulation of successful policies: if a hospital pursues the goal of increasing its output production efficiency, then such a goal can be accomplished by adopting the strategy of a better-performing peer. However, the adoption of a strategy of a better-performing hospital may not work in the German hospital market since not all hospitals represent a homogenous group. As the results of our clustering show, not every better-performing hospital is a better-performing peer for any other hospital. Nevertheless, we acknowledge this research is not without limitations. While clustering has been used to determine heterogeneity, it remains unclear what exactly constitutes heterogeneity. As heterogeneity is a relative concept that often requires intimate knowledge of the problem domain, this issue falls outside the scope of this study. The proposed framework can therefore be explored further in future research to examine the sources of heterogeneity, such as the differences in hospital environments.

References

Almeida Botega, L. de, Andrade, M. Viegas, & Guedes, G. Ramalho (2020). Brazilian hospitals' performance: an assessment of the unified health system (SUS). *Health Care Management Science*, *23*(3): 443–452.

Araújo, C., Barros, C. P., & Wanke, P. (2014). Efficiency determinants and capacity issues in Brazilian for-profit hospitals. *Health Care Management Science, 17*(2): 126–138.

Athanassopoulos, A. D., & Curram, S. P. (1996). A Comparison of Data Envelopment Analysis and Artificial Neural Networks as Tools for Assessing the Efficiency of Decision Making Units. *Journal of the Operational Research Society*, *47*(8): 1000–1016.

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., 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.

Bojnec, Š., & Latruffe, L. (2008). Measures of farm business efficiency. *Industrial Management & Data Systems, 108*(2): 258–270.

Brown, R. (2006). Mismanagement or mismeasurement? Pitfalls and protocols for DEA studies in the financial services sector. *European Journal of Operational Research, 174*(2): 1100–1116.

Charnes, A., Cooper, W. W., & Rhodes, E. (1978). Measuring the efficiency of decision making units. *European Journal of Operational Research, 2*(6): 429–444.

Cooper, W. W., Seiford, L. M., & Zhu, J. (Eds.) (2004). Handbook on Data Envelopment Analysis. Boston: Kluwer Academic Publishers (International Series in Operations Research & Management Science).

Cooper, W. W., Seiford, L. M., & Zhu, J. (2011). Handbook on data envelopment analysis. 2nd ed. New York: Springer (International series in operations research & management science, 0884-8289, 164). Available online at http://www.springer.com/gb/ BLDSS.

Daraio, C., & Simar, L. (Eds.) (2007). Advanced Robust and Nonparametric Methods in Efficiency Analysis: Methodology and Applications. Boston, MA: Springer US.

Dyson, R. G., Allen, R., Camanho, A. S., Podinovski, V. V., Sarrico, C. S., & Shale, E. A. (2001). Pitfalls and protocols in DEA. *European Journal of Operational Research, 132*(2): 245–259.

Emrouznejad, A., & Shale, E. (2009). A combined neural network and DEA for measuring efficiency of large scale datasets. *Computers & Industrial Engineering, 56*(1): 249–254.

Haas, D. A., & Murphy, F. H. (2003). Compensating for non-homogeneity in decision-making units in data envelopment analysis. *European Journal of Operational Research, 144*(3): 530–544.

Herr, A. (2008). Cost and technical efficiency of German hospitals: does ownership matter? *Health Economics, 17*(9): 1057–1071.

Herrera-Restrepo, O., Triantis, K., Seaver, W. L., Paradi, J. C., & Zhu, H. (2016). Bank branch operational performance: A robust multivariate and clustering approach. *Expert Systems with Applications, 50*: 107–119.

Hoff, A. (2007). Second stage DEA: Comparison of approaches for modelling the DEA score. *European Journal of Operational Research, 181*(1): 425–435.

Hudson, I. L., Keatley, M. R., & Lee, S. Y. (2011). Using Self-Organising Maps (SOMs) to assess synchronies: an application to historical eucalypt flowering records. *International journal of biometeorology, 55*(6): 879–904.

Jacobs, R., Smith, P. C., & Street, A. (2006). Measuring Efficiency in Health Care. Cambridge: Cambridge University Press.

Kingma, D. P., & Ba, J. (2014). Adam: A Method for Stochastic Optimization. Available online at http://arxiv.org/pdf/1412.6980v9.

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.

Kwon, H.-B. (2017). Exploring the predictive potential of artificial neural networks in conjunction with DEA in railroad performance modeling. *International Journal of Production Economics, 183*: 159–170.

Łukasik, S., Kowalski, P. A., Charytanowicz, M., & Kulczycki, P. (Eds.) (2016). Clustering using flower pollination algorithm and Calinski-Harabasz index: IEEE.

Mahmoudi, R., Emrouznejad, A., Khosroshahi, H., Khashei, M., & Rajabi, P. (2019). Performance evaluation of thermal power plants considering CO2 emission: A multistage PCA, clustering, game theory and data envelopment analysis. *Journal of Cleaner Production, 223*: 641–650.

Mann, H. B., & Whitney, D. R. (1947). On a Test of Whether one of Two Random Variables is Stochastically Larger than the Other. *The Annals of Mathematical Statistics, 18*(1): 50–60.

Mitropoulos, P., Mastrogiannis, N., & Mitropoulos, I. (2014). Seeking interactions between patient satisfaction and efficiency in primary healthcare: cluster and DEA analysis. *International Journal of Multicriteria Decision Making, 4*(3): p. 234.

Nedelea, I. Cristian, & Fannin, J. Matthew (2013). Technical efficiency of Critical Access Hospitals: an application of the two-stage approach with double bootstrap. *Health Care Management Science, 16*(1): 27–36.

Omrani, H., Shafaat, K., & Emrouznejad, A. (2018). An integrated fuzzy clustering cooperative game data envelopment analysis model with application in hospital efficiency. *Expert Systems with Applications, 114*: 615–628.

Ozcan, Y. A. (Ed.) (2014). Health Care Benchmarking and Performance Evaluation: An Assessment using Data Envelopment Analysis (DEA). Boston, MA: Springer US.

Pendharkar, P. C. (2005). A data envelopment analysis-based approach for data preprocessing. *IEEE Transactions on Knowledge and Data Engineering, 17*(10): 1379–1388.

Pendharkar, P. C. (2011). A hybrid radial basis function and data envelopment analysis neural network for classification. *Computers & Operations Research, 38*(1): 256–266.

Rocci, R., & Vichi, M. (2008). Two-mode multi-partitioning. *Computational Statistics & Data Analysis, 52*(4): 1984–2003.

Samoilenko, S., & Osei-Bryson, K.-M. (2008). Increasing the discriminatory power of DEA in the presence of the sample heterogeneity with cluster analysis and decision trees. *Expert Systems with Applications, 34*(2): 1568–1581.

Samoilenko, S., & Osei-Bryson, K.-M. (2010). Determining sources of relative inefficiency in heterogeneous samples: Methodology using Cluster Analysis, DEA and Neural Networks. *European Journal of Operational Research, 206*(2): 479–487.

Santín, D., Delgado, F. J., & Valiño, A. (2004). The measurement of technical efficiency: a neural network approach. *Applied Economics, 36*(6): 627–635.

Saxena, A., Prasad, M., Gupta, A., Bharill, N., Patel, O. Prakash, Tiwari, A. et al. (2017). A review of clustering techniques and developments. *Neurocomputing, 267*: 664–681.

Schneider, A. Maren, Oppel, E.-M., & Schreyögg, J. (2020). Investigating the link between medical urgency and hospital efficiency – Insights from the German hospital market. *Health Care Management Science, 23*(4): 649–660.

Simar, L., & Wilson, P. W. (1998). Sensitivity Analysis of Efficiency Scores: How to Bootstrap in Nonparametric Frontier Models. *Management Science, 44*(1): 49–61.

Simar, L., & Wilson, P. W. (2004). Performance of the Bootstrap for Dea Estimators and Iterating the Principle. In William W. Cooper, Lawrence M. Seiford, Joe Zhu (Eds.): Handbook on Data Envelopment Analysis, vol. 71. Boston: Kluwer Academic Publishers (International Series in Operations Research & Management Science): 265–298.

Simar, L., & Wilson, P. W. (2007). Estimation and inference in two-stage, semi-parametric models of production processes. *Journal of Econometrics, 136*(1): 31–64.

Tiemann, O., Schreyögg, J., & Busse, R. (2012). Hospital ownership and efficiency: A review of studies with particular focus on Germany. *Health Policy, 104*(2): 163–171.

Tone, K. (2001). A slacks-based measure of efficiency in data envelopment analysis. *European Journal of Operational Research, 130*(3): 498–509.

Tone, K. (2017). Advances in DEA Theory and Applications: With Extensions to Forecasting Models: Wiley. Available online at https://books.google.de/books?id=aU-wDgAAQBAJ.

Ünlü, R., & Xanthopoulos, P. (2019). Estimating the number of clusters in a dataset via consensus clustering. *Expert Systems with Applications, 125*: 33–39.

Weisberg, H. (1992). Central Tendency and Variability. Thousand Oaks, California. Available online at https://methods.sagepub.com/book/central-tendency-and-variability.

Wojcik, V., Dyckhoff, H., & Clermont, M. (2019). Is data envelopment analysis a suitable tool for performance measurement and benchmarking in non-production contexts? *Business Research, 12*(2): 559–595.

Appendix A. SOM function

In Figure I, we present the function developed and used for clustering which is based on the SOM-ANN. The function is developed by using *Scikit-learn* (https://scikit-learn.org/) which is an open-source platform for machine learning. However, the main codes can be also provided upon request.

***Sklearn* is an open-source machine learning platform**

```
from sklearn_som.som import SOM
from sklearn import metrics
import numpy as np
def CA_SOM(Data, i, j):
    # Data = {Beds, Physicians, Nurses, Inpatients, Outpatients, Surgeries}
    # i and j are the vertical and horizontal dimensions of the SOM, respectively.
    clusters = {}; CalinskiHarabasz = {}; Silhouette = {}; DaviesBouldin = {}
    # Create SOM and train: i and j can be adjusted in a loop, for example.
    SOMCluster = SOM(m=i, n=j, dim=6, lr=0.9, sigma=1.0, max_iter=2000)
    SOMCluster.fit(train_data, epochs=1, shuffle=True)
    clusters = SOMCluster.predict(train_data)
    # Calculate the metricts for SOM clusters
    CalinskiHarabasz = metrics.calinski_harabasz_score(train_data, clusters)
    Silhouette = metrics.silhouette_score(Data, clusters)
    DaviesBouldin = metrics.davies_bouldin_score(Data, clusters)
    # Concatenate the quility metrics
    metrics = {CalinskiHarabasz, Silhouette, DaviesBouldin}
    return clusters, metrics
# Import Bed_Cluster and Ownership_Cluster
# Calculate the metricts for clusters based on Bed size
CalinskiHarabasz_BedSize = metrics.calinski_harabasz_score(Data, Bed_Cluster)
Silhouette_BedSize = metrics.silhouette_score(Data, Bed_Cluster)
DaviesBouldin_BedSize = metrics.davies_bouldin_score(Data, Bed_Cluster)
# Calculate the metricts for clusters based on Owernership type
CalinskiHarabasz_Ownership = metrics.calinski_harabasz_score(Data, Ownership_Cluster)
Silhouette_Ownership = metrics.silhouette_score(Data, Ownership_Cluster)
DaviesBouldin_Ownership = metrics.davies_bouldin_score(Data, Ownership_Cluster)
```

Figure I. Function developed for clustering based on the SOM-ANN

Appendix B. Input-oriented SBM DEA model under VRS

We have a set of hospitals in each cluster as $DMU_j\ \forall j \in N = \{1,2,\dots,n\}$, each hospital having m inputs $\boldsymbol{X} = \left(x_{1j}, x_{2j}, \dots, x_{mj}\right)$ and s outputs $\boldsymbol{Y} = \left(y_{1j}, y_{2j} \dots, y_{rj}\right)$. The linear input-oriented SBM model under the VRS assumption can be written as Model (1).

$$\min \rho_h = 1 - \frac{1}{m}\sum_{i=1}^{m} \frac{s_i^-}{x_{ih}} \tag{1.1}$$

$$\text{s.t.} \quad x_{ih} = \sum_{j=1}^{n} x_{ij}\lambda_j + s_i^-, \forall i = 1, \dots, m \tag{1.2}$$

$$y_{rh} = \sum_{j=1}^{n} y_{rj}\lambda_j - s_r^+, \forall r = 1, \dots, s \tag{1.3}$$

$$\sum_{j=1}^{n} \lambda_j = 1 \quad (1.4)$$

$$\boldsymbol{s}^-, \boldsymbol{s}^+, \boldsymbol{\lambda} \geq \mathbf{0}, t > 0 \quad (1.5)$$

where ρ_h is the SBM-efficiency score of DMU_h. $\boldsymbol{s}^-$ and $\boldsymbol{s}^+$ are the vector of input and output slacks, respectively. $\boldsymbol{\lambda}$ is a non-negative vector and can modify the production possibility set by imposing some constraints on it, such as the VRS constraint $\sum_{j=1}^{n} \lambda_j = 1$. The optimal solution of the SBM DEA model can be defined as $\{\rho_h^*, \boldsymbol{\lambda}^*, \boldsymbol{s}^{-*}, \boldsymbol{s}^{+*}\}$. Figure II presents the function developed for solving Model (1) using *Gurobi Optimizer* (more information available at: https://www.gurobi.com/) in Python 3.8.

Gurobi Optimizer: Mathematical programming solver

```
def SBM_IO_VRS(X, Y):
    # Tone (2001) \ European Journal of Operational Research 130, 498–509.
    # Get number of DMUs (n), inputs (m) and outputs (s)
    n = len(X)
    m = len(X[0])
    s = len(Y[0])
    # Create arrays for saving the results
    Eff = {}; sol_Sm = {} ; sol_Sp = {} ; sol_lam = {}
    sol = {}
    # # Main loop over No. of DMUs
    for h in range(n):
        SBM_IO = Model("SBM-IO-VRS")
        # Variables
        Sm = SBM_IO.addVars(m, name="InputSlack")
        Sp = SBM_IO.addVars(s, name="OutputSlack")
        lam = SBM_IO.addVars(n, name="Lambda")
        # Constraints
        SBM_IO.addConstrs((sum(X[j][i]*lam[j] for j in range(n)) + Sm[i] == X[h][i] for i in range(m))) # 'CT1.2'
        SBM_IO.addConstrs((sum(Y[j][r]*lam[j] for j in range(n)) - Sp[r] == Y[h][r] for r in range(s))) # 'CT1.3'
        SBM_IO.addConstr((sum(lam[j] for j in range(n)) == 1)) # 'CT1.4'
        # Objective function
        SBM_IO.setObjective((1 - (1/m)*sum(Sm[i]/X[h][i] for i in range(m))), GRB.MINIMIZE)
        SBM_IO.optimize()
        if SBM_IO.status == GRB.INF_OR_UNBD:
            # Turn presolve off to determine whether model is infeasible
            # or unbounded
            SBM_IO.setParam(GRB.Param.Presolve, 0)
            SBM_IO.optimize()
        if SBM_IO.status == GRB.OPTIMAL:
            Eff[h] = SBM_IO.objVal
            print(f'DMU[{h+1}]: Optimal objective: {SBM_IO.objVal}')
            # SBM_IO.write('SBM_IO.sol')
            sol[h] = [(v.varName, v.X) for v in SBM_IO.getVars()]
            #sys.exit(0)
        elif SBM_IO.status != GRB.INFEASIBLE:
            print(f'DMU[{h+1}]: Optimization was stopped with status {SBM_IO.status}.')
            #sys.exit(0)
    return Eff, sol
```

Figure II. Function developed for solving SBM DEA model

Definition 1. (Projection). *The projection of* $DMU_o = (\boldsymbol{x}_o, \boldsymbol{y}_o)$ *onto the efficient frontiers can be defined by an optimal solution of the input-oriented SBM DEA model as* Eq. (2) (Tone, 2001, 2017).

$$\left(\boldsymbol{x}_o^p, \boldsymbol{y}_o^p\right) = (\boldsymbol{x}_h - \boldsymbol{s}^{-*}, \boldsymbol{y}_h + \boldsymbol{s}^{+*}) \quad (2)$$

The projected $DMU_h^p = \left(\boldsymbol{x}_h^p, \boldsymbol{y}_h^p\right)$ is SBM-input-efficient (Tone 2001). We use the SBM DEA model to compute efficiency scores for each hospital in the second stage of our proposed framework, relative efficiency analysis. Following this, the framework generates projections of the efficiency requirements for each inefficient hospital to become efficient.

Appendix C. Efficiency comparison of two hospital groups

The algorithm that is developed for efficiency comparison of two DMU groups (G_1 and G_2):

Step 1: Calculate the skewness of inefficiencies of both groups.

Step 2: If the inefficiencies are not skewed (symmetrically distributed), conduct the efficiency comparison based on the mean values. A parametric test such as the unpaired Student's t-test might be appropriate (Banker et al. 2010).

Step 3: If the inefficiencies are either positively or negatively skewed (asymmetrically distributed), the following are the procedures for testing the null hypothesis of a difference in efficiency between G_1 and G_2:

Step 3.1: Determine whether inefficiencies in G_1 and G_2 exhibit exponential distributions by using the Quantile-Quantile (Q-Q) plots. If so, the test statistic is therefore calculated as $\left(\sum_{j\in G_1}\rho_j^*/\|G_1\|\right)/\left(\sum_{j'\in G_2}\rho_{j'}^*/\|G_2\|\right)$ and assessed to the critical value of the F distribution with $(2\cdot\|G_1\|, 2\cdot\|G_2\|)$ degrees of freedom under the null hypothesis that there is no difference between them (Banker 1993).

Step 3.2: Determine whether inefficiencies in G_1 and G_2 exhibit half-normal distributions by using Q-Q plots. If so, the test statistic is therefore calculated as $\left(\sum_{j\in G_1}\left(\rho_j^*\right)^2/\|G_1\|\right)/\left(\sum_{j'\in G_2}\left(\rho_{j'}^*\right)^2/\|G_2\|\right)$ and assessed to the critical value of the F distribution with $(\|G_1\|, \|G_2\|)$ degrees of freedom under the null hypothesis that there is no difference between them.

Step 3.3: In the absence of such assumptions in steps 3.1 and 3.2, use a non-parametric test, such as Kolmogorov–Smirnov or Mann–Whitney tests. The results of the study conducted by Banker et al. (2010) indicate that the Mann–Whitney test performs better than Kolmogorov–Smirnov concerning error types I and II. Next, run the Mann–Whitney test to determine whether one of the random variables is stochastically greater than the other. In a combined and ordered sample of G_1 and G_2, the Mann–Whitney statistic is calculated by counting how many times each $\rho_j^*, j\in G_1$ occurs before $\rho_{j'}^*, j'\in G_2$. Define the random variable as Eq. (3).

$$D_{jj'}=\begin{cases}1 & \rho_j^*<\rho_{j'}^*\\ 0 & otherwise\end{cases} \tag{3}$$

Then, Mann–Whitney's statistic is given by $U=\sum_{j\in G_1}\sum_{j'\in G_2}D_{jj'}$. Mann and Whitney (1947) prove that for large samples of G_1 and G_2 ($\|G_1\|\ and\ \|G_2\|\geq 30$), Mann–Whitney's statistic is normally distributed with the mean of $\mu=\|G_1\|\cdot\|G_2\|/2$ and the variance of $\sigma^2=\left(\|G_1\|\cdot\|G_2\|\cdot(\|G_1\|+\|G_2\|+1)\right)/12$. Therefore, the large-sample (more than 20

observations) Mann–Whitney's test statistics can be approximated via $z = (U - \mu)/\sigma$ which follows a normal standard distribution function. Note, since there are a number of ties (i.e., the ranks of efficient DMUs) in each cluster, we need to revise the variance as $\sigma^2_{revised} = \sigma^2 \cdot (1 - \sum_t f_t^3 - f_t / f_t^3 - f_t)$, where t varies over the set of tied ranks and f_t represents frequency of the rank t. A further complication is that since we approximate a discrete distribution via a continuous one it is desirable to apply a continuity correction on the z-score as $z_{corrected} = U - \mu - 0.5 \cdot sign(U - \mu)/\sigma$.

Appendix D. Data preprocessing

In this study, the proposed framework is examined in the context of a large dataset of hospitals that were originally classified by the Federal Joint Committee (G-BA) in Germany in 2017. Data protection regulations prevent the dataset from being publicly available. Nevertheless, G-BA would send a copy to researchers upon official request (more information: https://www.g-ba.de/english/). In the German healthcare system, the G-BA, founded on 01.01.2004 due to the Healthcare Modernization Act, is the highest decision-making body. They establish guidelines that determine which medical treatments approximately 73 million insured people can claim. Furthermore, the G-BA establishes quality assurance measures for hospitals and healthcare practices. It is their responsibility to properly implement quality-improving measures. The implementation of individual quality assurance measures should be delegated as part of this overall responsibility. For the reporting year 2017, raw data includes all hospital quality reports from hospitals, the State Office for Quality Assurance, and the Institute for Quality Assurance and Transparency in Health Care at the end of medical transcription (MT). The preprocessing steps applied to the dataset in this study are illustrated in Figure III. Our dataset covers the following periods:

- Hospitals MT periods: October 15th to November 15th, 2018, and November 23rd to December 15th, 2018,
- State Office for Quality Assurance and Institute for Quality Assurance and Transparency in Health Care MT periods: November 15th to December 15th, 2018, and
- the subsequent reports of the State Office for Quality Assurance and the Institute for Quality Assurance and Transparency in Health Care occurring from January 20th to 23rd, 2019.

Appendix E. Quality criteria for clustering approaches

Figure IV shows the three quality criteria - CH-index, Silhouettes, and Davies-Bouldin - calculated to assess the homogeneity within hospitals clusters and the heterogeneity between clusters. These criteria are calculated in the function developed for the SOM (Figure I). Clusters that are dense and well separated achieve a high score on the CH-index. A clustering score of -1 is assigned for incorrect

clustering, whereas a clustering score of +1 is assigned to dense and well-separated clustering. Davies-Bouldin with a value close to zero indicates a more effective partition. Results show that cluster [1,3] outperforms other clusters on all three quality criteria.

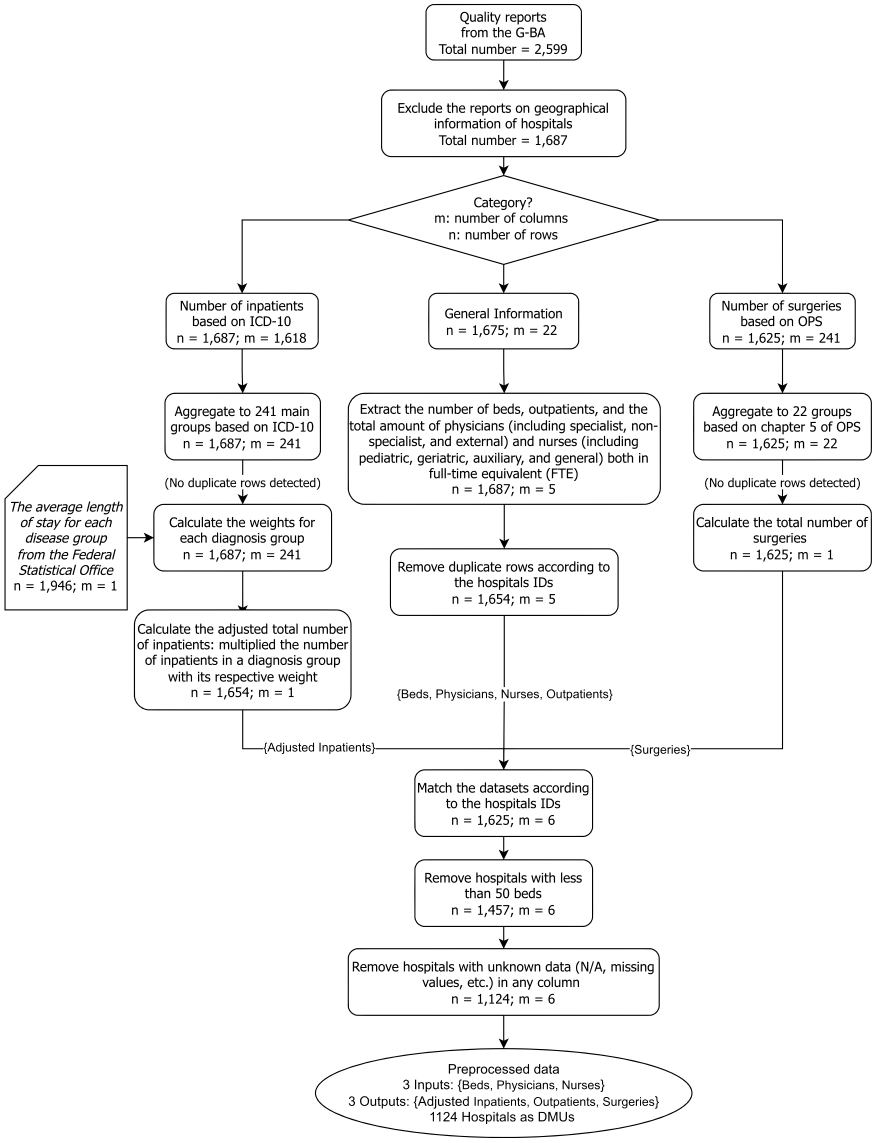


Figure III. Data preprocessing steps

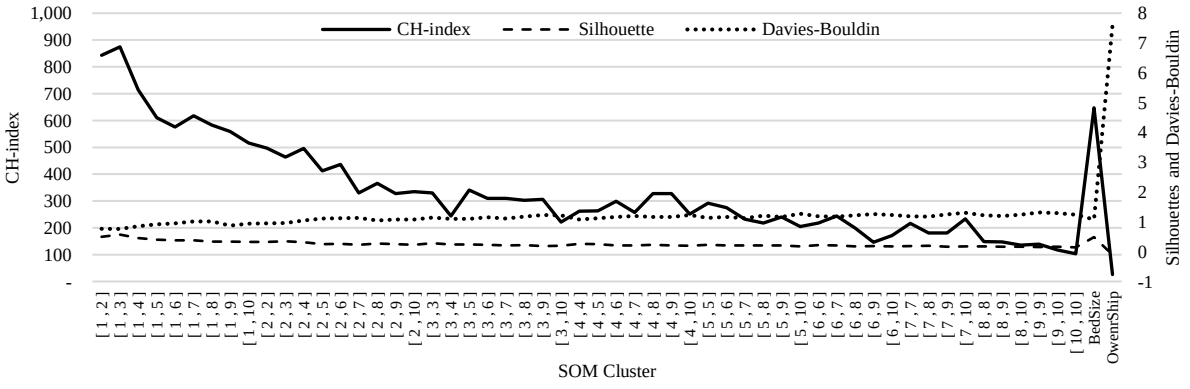


Figure IV. Quality criteria of clusters

Appendix F. Developed MLP-ANNs for creating TCM and BPM

As shown in Figure V, we have developed functions to create the TCMs and BPMs respectively by using two open-source platforms for machine learning: *TensorFlow* (more information available at: https://www.tensorflow.org/) and *Scikit-learn* (more information available at: https://scikit-learn.org/).

***TensorFlow* and *scikit-learn* are open-source machine learning platforms.**

```
from sklearn.model_selection import train_test_split
from sklearn import metrics
from tensorflow.keras.models import Sequential
from tensorflow.keras.layers import Dense
from tensorflow.keras.optimizers import Adam

def Create_TCM(Inputs, Outputs):
        in_dim = Inputs.shape[1]  # Input dimension
        out_dim = Output.shape[1] # Target dimension
        # Split train/test data
        xtrain, xtest, ytrain, ytest=train_test_split(Inputs, Output, test_size=0.15) # TCM_1 -> test_size=20
        print(\txtrain:', len(xtrain), 'ytrian:', len(ytrain))
        # Create the network
        TCM= Sequential()
        TCM.add(Dense(L1_k, input_dim=in_dim, activation='sigmoid')) # [L1_1, L1_2, L1_3] = [20, 20, 20]
        TCM.add(Dense(L2_k, input_dim=in_dim, activation='sigmoid')) # [L2_1, L2_2, L2_3] = [10, 10, 10]
        TCM.add(Dense(L3_k, activation='relu')) # [L3_1, L3_2, L3_3] = [10, 10, 10]
        TCM.add(Dense(out_dim))
        TCM.compile(loss='mape', optimizer='adam')
        # Set optimizer parameters
        keras.optimizers.Adam(lr=0.001, beta_1=0.9, beta_2=0.999, epsilon=None, decay=0.0, amsgrad=False)
        TCM.summary()
        ## Training
        TCM.fit(xtrain, ytrain, epochs=2000, batch_size=10, verbose=0,
                validation_split=0.05,validation_data=None)
        ypred = TCM.predict(xtest)
        print("\tTest MAPE: %.3f" % metrics.mean_absolute_percentage_error(ytest, ypred))
        return TCM

def Create_BPM(Data, Eff):
        in_dim = Data.shape[1]  # Input dimension
        out_dim = Eff.shape[1]  # Target dimension =1
        # Split train/test data
        xtrain, xtest, ytrain, ytest=train_test_split(Inputs, Output, test_size=0.15) # BPM_1 -> test_size=20
        print(\txtrain:', len(xtrain), 'ytrian:', len(ytrain))
        # Create the network
        BPM= Sequential()
        BPM.add(Dense(L1_k, input_dim=in_dim, activation='sigmoid')) # [L1_1, L1_2, L1_3] = [8, 10, 10]
        BPM.add(Dense(L2_k, activation='relu')) # [L2_1, L2_2, L2_3] = [8, 10, 10]
        BPM.add(Dense(out_dim))
        BPM.compile(loss='mape', optimizer='adam')
        # Set optimizer parameters
        keras.optimizers.Adam(lr=0.001, beta_1=0.9, beta_2=0.999, epsilon=None, decay=0.0, amsgrad=False)
        BPM.summary()
        ## Training
        BPM.fit(xtrain, ytrain, epochs=1500, batch_size=10, verbose=0, validation_split=0.05,validation_data=None)
        ypred = BPM.predict(xtest)
        print("\tTest MAPE: %.3f" % metrics.mean_absolute_percentage_error(ytest, ypred))
        return BPM
```

Figure V. Functions developed for creating TCMs and BPMs

Appendix III. A Mixed-Integer Slacks-Based-Measure Data Envelopment Analysis for Classifying Inputs and Outputs of German University Hospitals

Mansour Zarrin

Chair of Health Care Operations / Health Information Management, Faculty of Business and Economics, University of Augsburg, Universitätsstraße 16, 86159 Augsburg, Germany

Status: Submitted to Health Care Management Science, Category A.

Abstract. Standard Data Envelopment Analysis (DEA) models consider continuous-valued and known input and output statuses for measures. This paper proposes an extended Slacks-Based-Measure (SBM) DEA model to accommodate flexible (a measure that can play the role of input and output) and integer measures simultaneously. A flexible measure's most appropriate role (designation) is determined by maximizing the technical efficiency of each unit. The main advantage of the proposed model is that all inputs, outputs, and flexible measures can be expressed in integer values without inflation of efficiency scores since they are directly calculated by modifying input and output inefficiencies. Furthermore, we illustrate and examine the application of the proposed models with 28 university hospitals in Germany. We investigate the differences and common properties of the proposed models with the literature to shed light on both teaching and general inefficiencies. Results of inefficiency decomposition indicate that "Third-party funding income" that university hospitals receive from the research-granting agencies dominates the other inefficiencies sources.

Keywords. Data Envelopment Analysis; Integer-valued Measures; Flexible Measures; University Hospitals

1 Introduction

Data Envelopment Analysis (DEA) is a nonparametric approach introduced by Charnes et al. (1978) to estimate the relative efficiency of a set of homogeneous Decision-Making Units (DMUs) where utilize similar inputs to generate similar outputs. This basic model (from now referred to as CCR) has come up with a fruitful area for efficiency evaluation. DEA models can be categorized as radial and non-radial. The CCR represents the radial models where they cope with relative changes of inputs and/or outputs factors so that, the efficiency score imitates the proportional maximum output (input) expansion (reduction) rate which is common to all outputs (inputs). However, in many practical applications, not all inputs/outputs operate proportionally. Consider the hospitals as an instance, we utilize beds, physicians, and nurses as inputs where they may not change proportionally. There might be several non-radial slacks left that play an imperative role in reporting managerial efficiency, but they are not taken into account in the radial model. The Slacks-Based Measure (SBM) approach, on the other hand, disregards proportional changes and evaluates efficiency considering the input excess and output shortfall (slacks) directly (Tone 2001). Both non-radial and radial DEA models have been well-documented from the theoretical perspective in the literature (Tone 2017). In addition to the theoretical development of DEA models, their application is significantly growing since it is well-known as a reliable methodology, e.g., for healthcare (Kohl et al. 2019), higher education (Villano and Tran 2018), transportation (Stefaniec et al. 2020), and production process (Kourtzidis et al. 2021). However, to our knowledge, most of the previous studies done in the field of teaching hospital performance assessment use the basic DEA models and pay no attention to two principal challenges that exist in the real-world situation: integer-valued amounts and flexible measures. In the following subsections, these two issues are adequately addressed.

1.1 *Integrality-constrained DEA*

Conventional DEA models consider that inputs and outputs are continuous values. However, we face many real situations in which one or some of the inputs/outputs are unavoidably integer values, for instance, the number of beds (as input) and outpatients (as output) in the hospital performance assessment. Usually the first step in DEA application, after identifying the list of inputs and outputs, is determining the suitable technology or the Production Possibility Set (PPS). These technologies are grouped as non-convex and convex. The non-convex Free Disposal Hull (FDH) (Tulkens 2006) and the convex Constant and Variable Returns to Scale (CRS and VRS respectively) technologies are the most common choices. FDH targets are always feasible when some of the inputs/outputs are integer-valued since they project the units whose efficiency is to be evaluated onto one of the existing DMUs. In contrast, the PPS in both CRS and VRS assumes feasible operating points that are a convex combination of evaluating units without essentially considering any integrality constraint of some inputs/outputs. While imposing the integrality constraints by rounding off the optimum solution of the large integer

values may have not a major effect on the optimality, it is not the case with small integer values where a few units less or more can make a significant difference in the optimality (Lozano and Villa 2007; Kuosmanen and Matin 2009; Du et al. 2012). Assuming the integer-valued inputs/outputs as continuous values and arbitrarily rounding up (or down) of them may easily cause infeasibility (i.e., an operation point out of the PPS) or to a dominated (inferior) operating unit as mentioned by (Kuosmanen et al. 2015). As illustrated by a single input single output example in Figure 1, DMUs B and C are inefficient and their reference set includes DMUs A and D. The input excess of DMUs B and C are 1.5 and 2.67. That means 1.5- and 2.67-units reduction in the input of DMU B and C, respectively, project them on the green marks B' and C' on the efficient frontier (blue dashed line). However, if they are integer-valued, arbitrarily rounding up the input excess of C to 3.0 causes infeasibility, in other words, the red mark C'' where is out of the PPS. As it is clear from the graph, an arbitrary rounding down the input excess of B to 1.0 (B'') does not approach the efficient frontier.

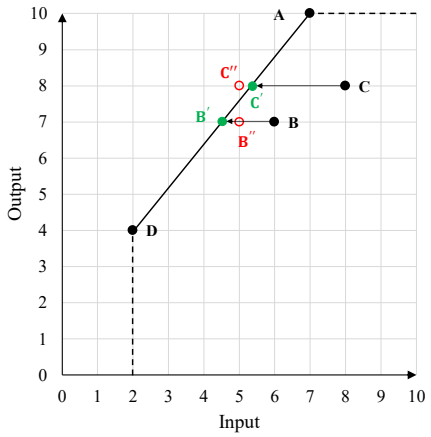


Figure 1. An example of infeasibility in the presence of integer-valued input under the VRS setting

1.2 *Flexible Measure DEA*

The usual setting for a DEA study is to evaluate DMUs, such as hospitals, according to specific input and output factors. The output represents the result of the DMU, while the input is intended to describe what led to the creation of that output. However, there exist some situations where some measures can play the role of either output or input. Consider, for example, the number of graduates or trainee nurses in a university hospital. These measures can constitute either input (two available human resources to the hospital) or output (trained staff, henceforth an advantage resulting from teaching/research funding). These measures are known as flexible or dual-role measures in DEA literature. In cases of ambiguity, it is imperative to adhere to the most equitable treatment possible for a particular DMU in order to decide the status of a variable. This ambiguity is further compounded if one views performance measurement from the perspective of an administrative organization as a manager. As with university hospitals, deciding whether graduates are to be regarded as an input or an output can have a tremendous impact on the funding received by each individual candidate. Therefore, these hospitals have a financial interest in using the least controversial and most fair method possible to assess efficiency.

The use of a factor as both an input and an output is not completely unheard of in the DEA framework. The status of flexible variables in DEA settings can be determined by at least three approaches. The first approach treats flexible measures on both the input and output sides simultaneously. For example, Beasley (1990) treats "research funding" on both the input and output sides at the same time in university efficiency measuring. Later, Cook et al. (2006) show that this treatment is not completely appropriate. Second, and perhaps most obvious, is to consider the issue from the standpoint of individual DMUs. A DEA model is run specifically for each DMU to determine the optimal role of each flexible measure. It could then be decided based on the majority choice among the DMUs what the overall input versus output status of any flexible measure is. In this case, it would seem to be the least controversial way to choose to apply a simple majority decision rule (Cook and Zhu 2007). As a third alternative, it is possible to consider the situation from the viewpoint of the manager of a collection of DMUs. Specifically, consider defining each flexible variable as an input or output so that the average or aggregate efficiency of the set of DMUs is maximized. An approach such as this would be useful if ties are encountered on a case-by-case basis (Cook and Zhu 2007; Cook et al. 2006; Ghiyasi and Cook 2021).

This study aims to develop an SBM DEA model that includes integer- and continuous-valued inputs, outputs, and flexible measures at the same time. Each flexible measure in the proposed model can be viewed as input, output, or both. The flexible measure's optimal role for the DMU being evaluated is dedicated to maximizing its technical efficiency. As a result, both the input surplus and output shortfall (slacks) may be present in the optimal solution set for each inefficient DMU. For efficient DMUs, flexible measures can be viewed both as input and output without affecting the degree of efficiency, since they are the ones with no slacks in their optimal solution. The proposed model has another advantage in that all three classes of measures can only take integer-valued slacks.

The rest of this paper is structured as follows. The literature on theoretical and application issues is reviewed in Section 2. In section 3, we review the advances in SBM DEA models in the presence of integer-valued and flexible measures. Then, we propose a new model as well as a new efficiency index. Section 4 presents the case study of the German university hospitals and the results of running the proposed models (efficiencies and slacks) and the developed ones in the literature. We also analyze the obtained results from the models and investigate the inefficiencies sources in this section. Finally, we wrap up our study and findings in Section 5.

2 Literature Review

This section provides an overview of the theoretical and application literature. We begin by examining studies related to the measurement of university hospital performance. Afterward, we review the theoretical development of the DEA models for integer-valued and flexible measures.

From a practical perspective, this study focuses on the performance evaluation of university hospitals. As reported in the health economics literature, teaching and university hospitals are more expensive than non-teaching counterparts (e.g., acute and general hospitals) since they engage in not only patient care but also in medical education and research. Therefore, this teaching/researching mission should be appropriately captured by defining proper measurements in the performance assessment process. One of the first studies in this field is conducted by Grosskopf et al. (2001). They compare the patient service provision of both non-teaching and teaching hospitals by the means of the basic DEA model. They apply the DEA model to a dataset that includes 556 non-teaching and 236 teaching hospitals in the US. Their results specify around 10% of the teaching hospitals can efficiently compare with non-teaching counterparts. Later, Grosskopf et al. (2004) evaluate the relative scale and technical efficiencies of 254 US teaching hospitals. They find that intensified competition results in superior efficiency deprived of cooperating teaching intensity. Ozcan et al. (2010) evaluate the performance of Brazilian teaching hospitals considering both medical care and teaching/research. They conclude their study by indicating the required changes for the inefficient teaching hospitals as some recommendations for public financing and teaching ratios. In another study, Lobo et al. (2014) study the efficiency of 104 teaching hospitals in Brazil. They use a two-stage weighted DEA model followed by logistic regression analysis in the second stage to examine the effect of non-discretionary variables (e.g., ownership type) on the efficiency scores. The result of the regression shows no significant relationship between ownership and efficiency. In the case of the German hospital market, recently, Schneider et al. (2020) conduct a study on efficiency analysis of German hospitals (including both teaching and non-teaching) with a focus on investigating the relation between medical urgency and efficiency. Their results show a lower efficiency for teaching hospitals compared to the non-teaching ones. This is because the same set of input/output with the non-teaching hospitals are only used in their DEA model and teaching function is not apprehended.

The integer DEA models have not attracted too much attention even though this situation can happen frequently in real-case applications. One reason for this may be the commitment of the DEA researchers to Linear Programming (LP) models since most LP DEA models can be proficiently solved even for big datasets using non-commercial solvers. To our knowledge, Lozano and Villa (2006) introduce the first DEA model whose inputs and outputs are intuitively constrained to take integer values only. They model their problem as a Mixed-Integer Linear Programming (MILP) for assessing efficiency of DMUs. Kuosmanen and Matin (2009) develop a new axiomatic foundation (namely, "natural disposability" and "natural divisibility") for DEA subject to the integrality constraints. They derive a new DEA PPS that fulfills the minimum extrapolation principle under their advanced axioms. They also present an MILP formula for assessing efficiency scores of Iranian university departments under the CRS assumption. Later, Kazemi Matin and Kuosmanen (2009) extend their axiomatic foundation for the integer DEA under VRS, non-increasing, and non-decreasing returns to scale.

Khezrimotlagh et al. (2013b) critique these two models and show that the input targets obtained from the model proposed by Kuosmanen and Matin (2009) and Kazemi Matin and Kuosmanen (2009) may not be less than those computed by the model developed by Lozano and Villa (2006). Jie et al. (2015) provide a technical note on the model proposed by Kuosmanen and Matin (2009) and improve their model into a rectified model. They show that the new model can effectively answer the problem of a counter case studied by Khezrimotlagh et al. (2013b). Since additive models target slacks directly in reporting the efficiency, they reveal higher discrimination power especially in the presence of integer values. Du et al. (2012) propose new models based on Andersen and Petersen's technique (Andersen and Petersen 1993) in which slacks are directly investigated in order to compute efficiency and super-efficiency scores when inputs and outputs are integer-valued.

For the purpose of incorporating flexible measures, Cook and Zhu (2007) present a modification of the standard CCR DEA model and illustrate its application in two practical problem settings. They develop their model using the MILP approach to suggest both a specific DMU model and an aggregate model as methods to originate the suitable descriptions for flexible measures. However, their technique may report incorrect inefficiency indices attributable to a computational problem as a result of utilizing a large positive number in their model. This situation is addressed by Toloo (2009). He revises Cook and Zhu's model so that it does not need to introduce a large positive number. The methodology classifies flexible measures either as input or output according to their contribution to technical efficiency optimization (optimum solution) based on MILP housing both possibilities simultaneously. Afterward, several studies try to propose further refinements (Toloo 2012; Toloo et al. 2021; Arana-Jiménez et al. 2020; Ghiyasi and Cook 2021). Some of the researchers also try to provide an ensuing and instructive discussion on the infeasibility issues of these models such as Amirteimoori and Emrouznejad (2012) and Sedighi Hassan Kiyadeh et al. (2019). The flexible SBM (FSBM) models have recently been addressed by some studies. Amirteimoori et al. (2013) introduce an FSBM for calculating the efficiency score where flexible measures are present. They show that if a DMU is perceived as efficient the flexible measure can play both input and output roles. Tohidi and Matroud (2017) develop an alternative non-oriented model to classify the status of flexible measures and determine returns to scale setting.

There are as well situations in the real world where certain measures can play either input or output roles and can only take integer values, for instance, the number of graduates. Such real situations result in new unified DEA models in which both integer-valued amounts and flexible measures are simultaneously addressed. Kordrostami et al. (2019) contribute to this topic by proposing an additive slacks-based approach which is also treatable under both VRS and CRS environments. However, additive models do not directly calculate the efficiency score of the DMU under evaluation in their objective function. Therefore, the final efficiency score can be (post) calculated using the SBM DEA model's definition. However, as pointed out by Khezrimotlagh et al. (2013a) the score of SBM for the

additive model may not result in an appropriate efficiency score. The SBM model measures the maximum possible slacks to minimize the efficiency score, whereas the additive model measures the maximum possible slacks without concerning the minimum efficiency score (Tone 2001). Therefore, the proposed model by Kordrostami et al. (2019) which is based on Du et al. (2012) may not report all inefficiencies (efficiency scores) correctly. Another issue that can be recognized is the way the flexible measures have been addressed in the FSBM models such as in Amirteimoori et al. (2013). They address the flexible measures in a way that deviates from the standard SBM. Since flexible measures can simultaneously be designated as input and output in the objective function, the averages in both the numerator (input excess) and denominator (output shortfall) are respectively computed using the fixed numbers of inputs plus the flexible measures and the fixed numbers of outputs plus the flexible measures regardless of the optimum solution where the status of the flexible measure is determined. Consequently, the efficiency score is overestimated compared to the efficiency score obtained from the standard SBM. Furthermore, the flexible measure may be differently classified for some DMUs. Boďa (2020) addresses this issue by proposing a modified model that distinguishes the same efficient and inefficient DMUs as Amirteimoori et al. (2013), however, realizes different projections for inefficient DMUs which means different classifications of flexible measures.

This study proposes an SBM model in which both flexible and integer measures are simultaneously presented. The main advantage of the proposed model is all input, output, and flexible measures can take integer-valued quantities without fluctuating the efficiency level. Furthermore, the technical efficiency score is directly calculated in the proposed SBM model and inflation of scores is prevented by modifying the input and output inefficiencies. The proposed model is developed based on the MILP approach then, can be easily solved by most non-commercial and open-source solvers. Furthermore, slack values of inputs, outputs, and flexible measures calculated by the proposed model are reported and compared with those obtained from Kordrostami et al. (2019). However, the same efficient and inefficient DMUs are detected as Kordrostami et al. (2019), the projections for inefficient DMUs and, consequently, classifications of flexible measures are different from each other. We also propose a new objective function for the model developed by Kordrostami et al. (2019) so that the new additive efficiency index falls between zero and one. The applicability of the introduced models is illustrated and scrutinized via a real-case dataset of German university hospitals. The main practical goal is to indicate the magnitude and source of inefficiencies for the university hospitals. This might support both local and national health authorities in decision-making processes including resource allocation, utilization, and planning.

3 Slacks-Based Measure Data Envelopment Analysis

We first present progress made in integer-valued and flexible DEA models in literature before moving on to the final proposed model. By doing this, readers should be able to better understand how the

models have evolved over the past two decades. Moreover, it allows us to point out how our proposed model advances other models by investigating their differences and commonalities. To begin with, it is worth mentioning the notations used in this paper as follows:

Sets and indices:

- N: set of DMUs, $N = \{1, \dots, n\}$
- I: set of real-valued inputs, $I = \{1, \dots, m\}$
- O: set of real-valued outputs, $O = \{1, \dots, s\}$
- K: set of real-valued flexible measures, $K = \{1, \dots, p\}$
- I^I: set of the integer-valued inputs, $I^I = \{1, \dots, m^I\}$
- I^{NI}: set of the non-integer valued inputs, $I^{NI} = \{1, \dots, m^{NI}\}$
- O^I: set of the integer-valued outputs, $O^I = \{1, \dots, s^I\}$
- O^{NI}: set of the non-integer-valued outputs, $O^{NI} = \{1, \dots, s^{NI}\}$
- K^I: set of integer-valued flexible measures, $K^I = \{1, \dots, p^I\}$
- K^{NI}: set of non-integer-valued flexible measures, $K^{NI} = \{1, \dots, p^{NI}\}$
- j: index of DMUs, $j \in N = \{1, \dots, n\}$
- i: index of inputs $i \in I = I^I \cup I^{NI}$
- r: index of outputs $r \in O = O^I \cup O^{NI}$
- k: index of flexible measures $k \in K = K^I \cup K^{NI}$

Parameters:

- x_{ij}: real-valued amounts of input i utilized by DMU_j
- y_{rj}: real-valued amounts of output r produced by DMU_j
- z_{kj}: real-valued amounts of flexible measure k utilized/produced by DMU_j

Decision variables:

- λ_j: coefficients of the convex linear combination
- s_i^x: real-valued amounts of input i excess
- s_r^y: real-valued amounts of output r shortfall
- s_{1k}^z, s_{2k}^z: real-valued amounts of flexible measure k slack designated as input and output, respectively
- $\tilde{s}_i^x$: integer-valued amounts of input i excess
- $\tilde{s}_r^y$: integer-valued amounts of output r shortfall
- $\tilde{s}_{1k}^z, \tilde{s}_{2k}^z$: integer-valued amounts of flexible measure k slack designated as input and output, respectively
- $\tilde{d}_k, d_k$: binary variables to indicate the role of integer- and non-integer valued flexible measure k, respectively

Auxiliary variables:

- $\tilde{x}_{ij}$: integer-valued reference point for input i utilized by DMU_j

- $\tilde{y}_{rj}$: integer-valued reference point for output r produced by DMU_j
- $\tilde{z}_{kj}$: integer-valued reference point for flexible measure k utilized/produced by DMU_j
- s'^{z}_{1k}, s'^{z}_{2k}: auxiliary variables for real flexible measure k as input and output, respectively
- δ_i^x: auxiliary variable for the real input i excess
- δ_r^y: auxiliary variable for the real output r shortfall
- $\delta_{1k}^z, \delta_{2k}^z$: auxiliary variables for real flexible measure k as input and output, respectively
- $\tilde{s}'^{z}_{1k}, \tilde{s}'^{z}_{2k}$: auxiliary variables for integer flexible measure k as input and output, respectively
- $\tilde{\delta}_i^x$: auxiliary variable for the integer input i excess
- $\tilde{\delta}_r^y$: auxiliary variable for the integer output r shortfall
- $\tilde{\delta}_{1k}^z, \tilde{\delta}_{2k}^z$: auxiliary variables for integer flexible measure k as input and output, respectively
- $a_{k'}, \tilde{a}_{k'}$: auxiliary binary decision variables

Now, assume we have n DMUs, $DMU_j\ \forall j = 1, \dots, n$, that utilize m inputs (real-valued inputs), $x_{ij}, \forall j, i = 1, \dots, m$ to produce s outputs (real-valued outputs) $y_{rj}, \forall j, r = 1, \dots, s$. The inputs and outputs can take only positive values[1] i.e., $\boldsymbol{x}, \boldsymbol{y} > \boldsymbol{0}$. Then, the SBM DEA model proposed by Tone (2001) can be formulated as:

$$[SBM]$$

$$\rho_h^{SBM} = Min\ \frac{1 - m^{-1}\left[\sum_{i \in I} \frac{s_i^x}{x_{ih}}\right]}{1 + s^{-1}\left[\sum_{r \in O} \frac{s_r^y}{y_{rh}}\right]} \tag{1.1}$$

$$s.t.\ \ x_{ih} = \sum_{j=1}^{n} \lambda_j x_{ij} + s_i^x,\ \ \forall i \in I \tag{1.2}$$

$$y_{rh} = \sum_{j=1}^{n} \lambda_j y_{rj} - s_r^y,\ \ \forall r \in O \tag{1.3}$$

$$\lambda_j, s_i^x, s_r^y \geq 0,\ \ \forall j, i, r \tag{1.4}$$

where ρ_h^{SBM} is the SBM efficiency score of the unit under evaluation DUM_h. $\boldsymbol{\lambda} = (\lambda_1, \dots, \lambda_n)$ is called the intensity vector which identifies the reference sets for DMU_h. $\boldsymbol{s}^x = (s_1^x, \dots, s_m^x)$ and $\boldsymbol{s}^y = \left(s_1^y, \dots, s_r^y\right)$ are respectively representing the input and output slacks. Note that Model (1) and the following models are formulated under the CRS setting, however, they can be reformulated under the VRS setting by simply adding $\sum_{j=1}^{n} \lambda_j = 1$ to the set of constraints.

3.1 *Integer-valued SBM DEA Model*

Suppose that some of the inputs and outputs are only valid in integer form. The input set $I = I^I \cup I^{NI}$, where I^I shows the index of the integer-valued inputs and I^{NI} shows the index of the rest of the inputs (non-integers). Similarly, the output set $O = O^I \cup O^{NI}$. To analyze the efficiency score of DMUs in the

[1]A difficulty arises with zero value measures since the slacks-based ratios are then divided by zero. To handle this problem Tone [Tone] provides some insights on how to deal with zeros in Section 6 of his paper.

presence of integer-valued quantities, Model (1) can be straightforwardly formulated based on the PPS defined by Du et al. (2012). Accordingly, the Integer-valued SBM (ISBM) DEA model can be written as follows:

$$[ISBM]$$

$$s.t. \quad \rho_h^{ISBM} = Min \ \frac{1-m^{-1}\left[\sum_{i\in I^{NI}}\frac{s_i^x}{x_{ih}}+\sum_{i\in I^I}\frac{\tilde{s}_i^x}{x_{ih}}\right]}{1+s^{-1}\left[\sum_{r\in O^{NI}}\frac{s_r^y}{y_{rh}}+\sum_{r\in O^I}\frac{\tilde{s}_r^y}{y_{rh}}\right]} \tag{2.1}$$

$$x_{ih} = \sum_{j=1}^{n}\lambda_j x_{ij} + s_i^x, \ \forall i \in I^{NI} \tag{2.2}$$

$$y_{rh} = \sum_{j=1}^{n}\lambda_j y_{rj} - s_r^y, \ \forall r \in O^{NI} \tag{2.3}$$

$$\tilde{x}_{ih} \geq \sum_{j=1}^{n}\lambda_j x_{ij}, \ \forall i \in I^I \tag{2.4}$$

$$\tilde{x}_{ih} = x_{ih} - \tilde{s}_i^x, \ \forall i \in I^I \tag{2.5}$$

$$\tilde{y}_{rh} \leq \sum_{j=1}^{n}\lambda_j y_{rj}, \ \forall r \in O^I \tag{2.6}$$

$$\tilde{y}_{rh} = y_{rh} - \tilde{s}_r^y, \ \forall r \in O^I \tag{2.7}$$

$$\lambda_j, s_i^x, s_r^y, \tilde{s}_i^x, \tilde{s}_r^y \geq 0, \ \forall j, i, r, k \tag{2.8}$$

$$\tilde{x}_{ih}, \tilde{y}_{rh} \ integer \ \forall i \in I^I, r \in O^I \tag{2.9}$$

where ρ_h^{ISBM} shows the efficiency score of DMU_h in the presence of integer measures. Du *et al*. [11] 's model does not offer a zero-to-one integrated efficiency score, as in the standard additive DEA model. Model (2) differs from the Du et al. (2012) model in its definition of the efficiency index (the objective function), which mirrors SBM's efficiency score (Tone 2001). Variables $\tilde{s}_i^x$ and $\tilde{s}_r^y$ are respectively non-radial slacks for integer-valued inputs and outputs while variables $\tilde{x}_{ih}$ and $\tilde{y}_{rh} \in \mathbb{Z}^+$are the integer-valued reference points (targets) for inputs and outputs of DMU_h, respectively. The slack variables $\tilde{s}_i^x$ and $\tilde{s}_r^y$ signify the absolute difference between the reference points ($\tilde{x}_{ih}$ and $\tilde{y}_{rh}$) and the integer-valued inputs and outputs. As shown in Figure 1, under the VRS setting, the integer DEA targets may not lie within the feasible area (the convex hull). That is why the modeling of the relationship between the convex linear combination and the integer-valued targets i.e., Eqs. (2.4) and (2.5) for integer-valued inputs, and Eqs. (2.6) and (2.7) for integer-valued outputs are slightly different from real-valued counterparts in Model (1). In other words, by defining the integer-valued reference points ($\tilde{x}_{ih}$ and $\tilde{y}_{rh}$) we guarantee the feasibility of the integer DEA model (Du et al. 2012).

3.2 *Flexible SBM DEA Model*

Consider p flexible measures shown by $z_{kj}, \forall j = \{1, \dots, n\}, k = \{1, \dots, p\}$ whose statuses (input or output) are unknown. To incorporate these measures, Model (1) can be reformulated based on the SBM model proposed by Amirteimoori et al. (2013) for classifying the flexible measures as follows:

$[FSBM]$

$$s.t. \quad \rho_h^{FSBM} = Min \ \frac{1-\left(m+\left(p-\sum_{k=1}^{p} d_k\right)\right)^{-1}\left[\sum_{i\in I}\frac{s_i^x}{x_{ih}}+\sum_{k\in K}\frac{s_{1k}^z}{z_{kh}}\right]}{1+\left(s+\sum_{k=1}^{p} d_k\right)^{-1}\left[\sum_{r\in O}\frac{s_r^y}{y_{rh}}+\sum_{k\in K}\frac{s_{2k}^z}{z_{kh}}\right]} \tag{3.1}$$

$$x_{ih} = \sum_{j=1}^{n}\lambda_j x_{ij} + s_i^x, \ \forall i \in I \tag{3.2}$$

$$y_{rh} = \sum_{j=1}^{n}\lambda_j y_{rj} - s_r^y, \ \forall r \in O \tag{3.3}$$

$$z_{kh} = \sum_{j=1}^{n}\lambda_j z_{kj} + s_{1k}^z - s_{2k}^z, \ \forall k \in K \tag{3.4}$$

$$s_{1k}^z \cdot s_{2k}^z = 0, \ \forall k \in K \tag{3.5}$$

$$\lambda_j, s_i^x, s_r^y, s_{1k}^z, s_{2k}^z \geq 0, \forall j, i, r, k \tag{3.6}$$

where s_{1k}^z and s_{2k}^z are the slacks vectors responding to the flexible measures treating as inputs and outputs, respectively. $s_{1k}^z > 0$ results in designating z_{ko} as input and $s_{2k}^z > 0$ means z_{ko} plays the role of output in the PPS. Since z_{kh} must be either designated as input or output, the unique status of it in the PPS is indicated by Eq. (3.5). The nonlinearity of this constraint can be handled by introducing a large positive number $\mathcal{M}$ and a binary decision variable $d_k, \forall k$ that assures one and only one of the variables s_{1k}^z and s_{2k}^z takes positive (non-zero) values simultaneously. Then, Eq. (3.5) can be replaced with the following equivalent linear constraints:

$$s_{1k}^z \leq \mathcal{M} \cdot (1 - d_k), \qquad \forall k \in K \tag{3.5.1}$$

$$s_{2k}^z \leq \mathcal{M} \cdot d_k, \qquad \forall k \in K \tag{3.5.2}$$

This condition should be reflected in the objection function Eq. (3.1) as well. In other words, if $s_{1k}^z > 0, \forall k$ ($d_k = 0, \forall k$) then $s_{2k}^z = 0$ ($d_k = 1, \forall k$) and the total number of inputs is $(m + p)$ in the numerator consequently, the total number of the outputs in the denominator is (s). However, this issue is skipped by Amirteimoori et al. (2013), and the number of inputs and outputs they utilize are $(m + p)$ and $(s + p)$, respectively. In other words, they consider the number of flexible measures at the same time in both numerator and denominator of the objective function. This results in overestimating the efficiency score since the second term of both numerator and denominator is underestimated. This issue can be solved by redefining the efficiency score as Eq (3.1). However, the objective function of Model (3) is non-linear. A linear counterpart of Model (3) is proposed by Boďa (2020). He modifies the FSBM model proposed by Amirteimoori et al. (2013) and proposes the following model:

$[mFSBM]$

$$\rho_h^{mFSBM} = Min \ \frac{1-\left[\sum_{i\in I}\frac{\delta_i^x}{x_{ih}}+\sum_{k\in K}\frac{\delta_{1k}^z}{z_{kh}}\right]}{1+\left[\sum_{r\in O}\frac{\delta_r^y}{y_{rh}}+\sum_{k\in K}\frac{\delta_{2k}^z}{z_{kh}}\right]} \tag{4.1}$$

$$s.t. \quad x_{ih} = \sum_{j=1}^{n}\lambda_j x_{ij} + s_i^x, \ \forall i \in I \tag{4.2}$$

$$y_{rh} = \sum_{j=1}^{n}\lambda_j y_{rj} - s_r^y, \ \forall r \in O \tag{4.3}$$

$$z_{kh} = \sum_{j=1}^{n} \lambda_j z_{kj} + s_{1k}^z - s_{2k}^z, \ \forall k \in K \tag{4.4}$$

$$s_{1k}^z \leq \mathcal{M} \cdot (1 - d_k), \ \forall k \in K \tag{4.5.1}$$

$$s_{2k}^z \leq \mathcal{M} \cdot d_k, \ \forall k \in K \tag{4.5.2}$$

$$\sum_{k'=0}^{p} k' \cdot \ a_{k'} = \sum_{k=1}^{p} d_k \tag{4.6}$$

$$\sum_{k'=0}^{p} a_{k'} = 1 \tag{4.7}$$

$$-(1 - a_{k'}) \cdot \mathcal{M} + \delta_i^x \cdot (m + p - k') \leq s_i^x \leq (1 - a_{k'}) \cdot \mathcal{M} + \delta_i^x \cdot (m + p - k'), \ \forall k', i \tag{4.8}$$

$$-(1 - a_{k'}) \cdot \mathcal{M} + \delta_r^y \cdot (s + k') \leq s_r^y \leq (1 - a_{k'}) \cdot \mathcal{M} + \delta_r^y \cdot (s + k'), \ \forall k', r \tag{4.9}$$

$$-(1 - a_{k'}) \cdot \mathcal{M} + \delta_{1k}^z \cdot (m + p - k') \leq s_{1k}^z \leq (1 - a_{k'}) \cdot \mathcal{M} + \delta_{1k}^z \cdot (m + p - k'), \ \forall k', k \tag{4.10}$$

$$-(1 - a_{k'}) \cdot \mathcal{M} + \delta_{2k}^z \cdot (s + k') \leq s_{2k}^z \leq (1 - a_{k'}) \cdot \mathcal{M} + \delta_{2k}^z \cdot (s + k'), \ \forall k', k \tag{4.11}$$

$$\lambda_j, s_i^x, s_r^y, s_{1k}^z, s_{2k}^z, s'^z_{1k}, s'^z_{2k}, \delta_i^x, \delta_r^y, \delta_{1k}^z, \delta_{2k}^z \geq 0, \ \forall j, i, r, k \tag{4.12}$$

$$d_k, a_{k'} \in \{0,1\}, \ \forall k, k' \tag{4.13}$$

where the optimal solution of the model determines the source of overestimating efficiency scores $\sum_{k=1}^{K} d_k$. This issue can be fixed by introducing an auxiliary binary variable $a_{k'}, \forall k' = \{0, \dots, p\}$ which controls the optimized number of flexible measures indicated as outputs $\sum_{k=1}^{p} d_k$. Constraints (4.6) to (4.13) ensure that the decision variables $\delta_i^x = \left(m + \left(p - \sum_{k=1}^{p} d_k\right)\right)^{-1} \cdot s_i^x, \forall i$, $\delta_r^y = \left(r + \sum_{k=1}^{p} d_k\right)^{-1} \cdot s_r^y, \forall r$, $\delta_{1k}^z = \left(m + \left(p - \sum_{k=1}^{p} d_k\right)\right)^{-1} \cdot s_{1k}^z, \forall k$, and $\delta_{2k}^z = \left(s + \sum_{k=1}^{p} d_k\right)^{-1} \cdot \delta_{2k}^z, \forall k$. Therefore, the efficiency score is calculated based on the correct total number of inputs and outputs. Eqs. (4.6) and (4.7) ensure the abovementioned equalities are accomplished only and only for $k' = \sum_{k=1}^{p} d_k$ (equivalently, $a_{\sum_{k=1}^{p} d_k} = 1$) in Constraints (4.8) to (4.11) otherwise, they turn into free limits. The conditions defined for flexible measures in Model (3) are valid in this model as well. Let $d_k = 1$ then $s_{1k}^z = 0, \forall k$, $s_{2k}^z > 0, \forall k$, and the flexible measure z_{ko} is designated as output. In contrast, if $d_k = 0$ then z_{ko} plays the role of input.

3.3 *Integer-valued Flexible SBM DEA Model*

In the presence of both integer and flexible measures ($K = K^I \cup K^{NI}$), Kordrostami et al. (2019) develop the additive model proposed by Du et al. (2012) to assess the relative efficiency. Our first step towards assessing the model's properties is to write the model as follows:

$$[FISBM]$$

$$\tau_h^{FISBM} = Max \ \sum_{i \in I^{NI}} \frac{s_i^x}{x_{ih}} + \sum_{i \in I^I} \frac{\tilde{s}_i^x}{x_{ih}} + \sum_{k \in K^{NI}} \frac{s_{1k}^z}{z_{ko}} + \sum_{k \in K^I} \frac{\tilde{s}_{1k}^z}{z_{ko}} + \sum_{r \in O^{NI}} \frac{s_r^y}{y_{ro}} + \sum_{r \in O^I} \frac{\tilde{s}_r^y}{y_{ro}} + \sum_{k \in K^{NI}} \frac{s_{2k}^z}{z_{ko}} + \sum_{k \in K^I} \frac{\tilde{s}_{2k}^z}{z_{ko}} \tag{5.1}$$

$$s.t. \quad x_{ih} = \sum_{j=1}^{n} \lambda_j x_{ij} + s_i^x, \ \forall i \in I^{NI} \tag{5.2}$$

$$y_{rh} = \sum_{j=1}^{n} \lambda_j y_{rj} - s_r^y, \ \forall r \in O^{NI} \quad (5.3)$$

$$z_{kh} = \sum_{j=1}^{n} \lambda_j z_{kj} + s_{1k}^z - s_{2k}^z, \ \forall k \in K^{NI} \quad (5.4)$$

$$s_{1k}^z \cdot s_{2k}^z = 0, \ \forall k \in K^{NI} \quad (5.5)$$

$$\tilde{x}_{ih} \geq \sum_{j=1}^{n} \lambda_j x_{ij}, \ \forall i \in I^I \quad (5.6)$$

$$\tilde{x}_{ih} = x_{ih} - \tilde{s}_i^x, \ \forall i \in I^I \quad (5.7)$$

$$\tilde{y}_{rh} \leq \sum_{j=1}^{n} \lambda_j y_{rj}, \ \forall r \in O^I \quad (5.8)$$

$$\tilde{y}_{rh} = y_{rh} - \tilde{s}_r^y, \ \forall r \in O^I \quad (5.9)$$

$$\tilde{z}_{kh} = \sum_{j=1}^{n} \lambda_j z_{kj} + \tilde{s}'^z_{1k} - \tilde{s}'^z_{2k}, \ \forall k \in K^I \quad (5.10)$$

$$\tilde{z}_{kh} = z_{kj} - \tilde{s}_{1k}^z + \tilde{s}_{2k}^z, \ \forall k \in K^I \quad (5.11)$$

$$\tilde{s}'^z_{1k} \cdot \tilde{s}'^z_{2k} = 0, \ \forall k \in K^I \quad (5.12)$$

$$\tilde{s}_{1k}^z \cdot \tilde{s}_{2k}^z = 0, \ \forall k \in K^I \quad (5.13)$$

$$\tilde{s}'^z_{1k} \cdot \tilde{s}_{2k}^z = 0, \ \forall k \in K^I \quad (5.14)$$

$$\tilde{s}_{1k}^z \cdot \tilde{s}'^z_{2k} = 0, \ \forall k \in K^I \quad (5.15)$$

$$\lambda_j, s_i^x, s_r^y, s_{1k}^z, s_{2k}^z, \tilde{s}_i^x, \tilde{s}_r^y, \tilde{s}_{1k}^z, \tilde{s}_{2k}^z \geq 0, \forall j, i, r, k \quad (5.16)$$

$$\tilde{x}_{ih}, \tilde{y}_{rh}, \tilde{z}_{kh} \ integer \ \forall i \in I^I, r \in O^I, k \in K^I \quad (5.17)$$

where τ_h^{FISBM} is the maximum summation of slacks. Similar to Model (2), a new integer decision variable $\tilde{z}_{ko}, \forall k$ is introduced which represents integer-valued projection points for flexible measure k of DMU_o. To calculate the efficiency score, Kordrostami et al. (2019) calculate the optimum value of slacks $\boldsymbol{s}^* = (\boldsymbol{s}^{*x}, \boldsymbol{s}^{*y}, \tilde{\boldsymbol{s}}^{*x}, \tilde{\boldsymbol{s}}^{*y}, \boldsymbol{s}^{*z}_1, \boldsymbol{s}^{*z}_2)$ and the determined status of flexible measure $\boldsymbol{d}^*$ obtained from Model (5). Then, they use the SBM's scalar measure as a posteriori efficiency index based on a set of optimal solution from Model (5) as follows:

$$\zeta_h^{*FISBM} = \frac{1-\left(m+\left(p-\sum_{k=1}^{p} d_k^*\right)\right)^{-1}\left[\sum_{i\in I^{NI}} \frac{s_i^{*x}}{x_{ih}} + \sum_{i\in I^I} \frac{\tilde{s}_i^{*x}}{x_{ih}} + \sum_{k\in K^{NI}} \frac{s_{1k}^{*z}}{z_{kh}} + \sum_{k\in K^I} \frac{\tilde{s}_{1k}^{*z}}{z_{kh}}\right]}{1+\left(s+\sum_{k=1}^{p} d_k^*\right)^{-1}\left[\sum_{r\in O^{NI}} \frac{s_r^{*y}}{y_{rh}} + \sum_{r\in O^I} \frac{\tilde{s}_r^{*y}}{y_{rh}} + \sum_{k\in K^{NI}} \frac{s_{2k}^{*z}}{z_{kh}} + \sum_{k\in K^I} \frac{\tilde{s}_{2k}^{*z}}{z_{kh}}\right]} \quad (6)$$

This model (as an additive model) deals directly with both integer- and real-valued input excesses and output shortfalls. However, it has no ratio efficiency term (scalar measure) per se. Model (5) is able to discriminate inefficient from efficient DMUs by looking for slacks, but it is unable to assess the real degree of inefficiency (Tone 2017; Khezrimotlagh et al. 2013a). Mathematically speaking, $\min\left[\frac{1-s^x/m}{1+s^y/s}\right] \not\equiv \max\left[\frac{s^x}{m} + \frac{s^y}{s}\right]$. For example, consider $s^x + s^y = 0.2 + 0.4 = 0.8$ which is greater than $s^x + s^y = 0.4 + 0.3 = 0.7$ but $\frac{1-s^x}{1+s^y} = \frac{1-0.2}{1+0.4} = 0.5$ is not less than $\frac{1-s^x}{1+s^y} = \frac{1-0.4}{1+0.3} = 0.46$. Therefore, we introduce a modified SBM DEA model (hereafter *mFISBM*) in an attempt to define the efficiency index directly based on the slacks and in the presence of integer and flexible measures as Model (7).

$[mFISBM]$

$$\rho_h^{mFISBM} = Min \frac{1-\left[\sum_{i\in I^{NI}}\frac{\delta_i^x}{x_{ih}}+\sum_{i\in I^{I}}\frac{\tilde{\delta}_i^x}{x_{ih}}+\sum_{k\in K^{NI}}\frac{\delta_{1k}^z}{z_{kh}}+\sum_{k\in K^{I}}\frac{\tilde{\delta}_{1k}^z}{z_{kh}}\right]}{1+\left[\sum_{r\in O^{NI}}\frac{\delta_r^y}{y_{rh}}+\sum_{i\in O^{I}}\frac{\tilde{\delta}_r^y}{y_{rh}}+\sum_{k\in K^{NI}}\frac{\delta_{2k}^z}{z_{kh}}+\sum_{k\in K^{I}}\frac{\tilde{\delta}_{2k}^z}{z_{kh}}\right]} \quad (7.1)$$

$s.t.$ (5.2) – (5.15)

$$\sum_{k'=0}^{p^{NI}} k' \cdot a_{k'} = \sum_{k=1}^{p^{NI}} d_k \quad (7.2)$$

$$\sum_{k'=0}^{p^{NI}} a_{k'} = 1 \quad (7.3)$$

$$-(1-a_{k'})\cdot\mathcal{M} + \delta_i^x \cdot (m^{NI}+p^{NI}-k') \le s_i^x \le (1-a_{k'})\cdot\mathcal{M} + \delta_i^x \cdot (m^{NI}+p^{NI}-k'),\ \forall i \in I^{NI}, k' = 0,\dots,p^{NI} \quad (7.4)$$

$$-(1-a_{k'})\cdot\mathcal{M} + \delta_r^y \cdot (s^{NI}+k') \le s_r^y \le (1-a_{k'})\cdot\mathcal{M} + \delta_r^y \cdot (s^{NI}+k'),\ \forall r \in O^{NI}, k' = 0,\dots,p^{NI} \quad (7.5)$$

$$-(1-a_{k'})\cdot\mathcal{M} + \delta_{1k}^z \cdot (m^{NI}+p^{NI}-k') \le s_{1k}^z \le (1-a_{k'})\cdot\mathcal{M} + \delta_{1k}^z \cdot (m^{NI}+p^{NI}-k'),\ \forall k \in K^{NI}, k' = 0,\dots,p^{NI} \quad (7.6)$$

$$-(1-a_{k'})\cdot\mathcal{M} + \delta_{2k}^z \cdot (s^{NI}+k') \le s_{2k}^z \le (1-a_{k'})\cdot\mathcal{M} + \delta_{2k}^z \cdot (s^{NI}+k'),\ \forall k \in K^{NI}, k' = 0,\dots,p^{NI} \quad (7.7)$$

$$\sum_{k'=0}^{p^{I}} k' \cdot \tilde{a}_{k'} = \sum_{k=1}^{p^{I}} \tilde{d}_k \quad (7.8)$$

$$\sum_{k'=0}^{p^{I}} \tilde{a}_{k'} = 1 \quad (7.9)$$

$$-(1-\tilde{a}_{k'})\cdot\mathcal{M} + \tilde{\delta}_i^x \cdot (m^{I}+p^{I}-k') \le \tilde{s}_i^x \le (1-\tilde{a}_{k'})\cdot\mathcal{M} + \tilde{\delta}_i^x \cdot (m^{I}+p^{I}-k'),\ \forall i \in I^{I}, k' = 0,\dots,p^{I} \quad (7.10)$$

$$-(1-\tilde{a}_{k'})\cdot\mathcal{M} + \tilde{\delta}_r^y \cdot (s^{I}+k') \le \tilde{s}_r^y \le (1-\tilde{a}_{k'})\cdot\mathcal{M} + \tilde{\delta}_r^y \cdot (s^{I}+k'),\ \forall r \in O^{I}, k' = 0,\dots,p^{I} \quad (7.11)$$

$$-(1-\tilde{a}_{k'})\cdot\mathcal{M} + \tilde{\delta}_{1k}^z \cdot (m^{I}+p^{I}-k') \le \tilde{s}_{1k}^z \le (1-\tilde{a}_{k'})\cdot\mathcal{M} + \tilde{\delta}_{1k}^z \cdot (m^{I}+p^{I}-k'),\ \forall k \in K^{I}, k' = 0,\dots,p^{I} \quad (7.12)$$

$$-(1-\tilde{a}_{k'})\cdot\mathcal{M} + \tilde{\delta}_{2k}^z \cdot (s^{I}+k') \le \tilde{s}_{2k}^z \le (1-\tilde{a}_{k'})\cdot\mathcal{M} + \tilde{\delta}_{2k}^z \cdot (s^{I}+k'),\ \forall k \in K^{I}, k' = 0,\dots,p^{I} \quad (7.13)$$

$$\lambda_j, s_i^x, s_r^y, s_{1k}^z, s_{2k}^z, s'^z_{1k}, s'^z_{2k}, \delta_i^x, \delta_r^y, \delta_{1k}^z, \delta_{2k}^z, \tilde{s}_i^x, \tilde{s}_r^y, \tilde{s}_{1k}^z, \tilde{s}_{2k}^z, \tilde{s}'^z_{1k}, \tilde{s}'^z_{2k}, \tilde{\delta}_i^x, \tilde{\delta}_r^y, \tilde{\delta}_{1k}^z, \tilde{\delta}_{2k}^z \ge 0, \forall j,i,r,k \quad (7.14)$$

$$\tilde{x}_{ih}, \tilde{y}_{rh}, \tilde{z}_{kh}\ integer\ \forall i \in I^I, r \in O^I, k \in K^I \quad (7.15)$$

$$d_k, a_{k'}, \tilde{d}_k, \tilde{a}_{k'} \in \{0,1\}, \forall k, k' \quad (7.16)$$

where similar to Model (4), the set of decision variables $\{\boldsymbol{\delta}^x, \boldsymbol{\delta}^y, \boldsymbol{\delta}^z\}$, and their equivalents for integer-valued measures (i.e., $\{\tilde{\boldsymbol{\delta}}^x, \tilde{\boldsymbol{\delta}}^y, \tilde{\boldsymbol{\delta}}^z\}$) make sure that the efficiency score is calculated based on the correct total number of inputs and outputs in the objective function by setting the boundaries of Constraints (7.4) to (7.7) and Constraints (7.10) to (7.13) via introducing the binary decision variable set $\{d_k, a_{k'}, \tilde{d}_k, \tilde{a}_{k'}\}$. The nonlinear Constraints (5.12) to (5.15) can be equivalently reformulated as the following set of linear constraints:

$$\tilde{s}_{1k}^z \le \mathcal{M} \cdot (1-\tilde{d}_k), \qquad \forall k \in K^I \quad (5.12.1)$$

$$\tilde{s}_{2k}^{z} \leq \mathcal{M} \cdot \tilde{d}_k, \qquad \forall k \in K^I \tag{5.13.1}$$

$$\tilde{s}'^{z}_{1k} \leq \mathcal{M} \cdot \left(1 - \tilde{d}_k\right), \qquad \forall k \in K^I \tag{5.14.1}$$

$$\tilde{s}'^{z}_{2k} \leq \mathcal{M} \cdot \tilde{d}_k, \qquad \forall k \in K^I \tag{5.15.1}$$

The modified flexible integer-valued SBM DEA model introduced here has all properties of the SBM DEA model originally developed by Tone (2001). The *mFISBM* is units-invariant, i.e., the value of $\rho^{*mFISBM}_{h}$ (≤ 1) is autonomous of the units in which the inputs, outputs, and flexible measures are assessed. It can as well be confirmed that ρ_h^{mFISBM} is monotone decreasing in all input excesses, output shortfalls, and flexible slacks. To such an extent, a larger value results larger performance score in the attainment of the efficient frontier/facet. $\rho^{*mFISBM}_{o} = 1$ means $\boldsymbol{s^{x^*}} = \mathbf{0}, \boldsymbol{s^{y^*}} = \mathbf{0}, \boldsymbol{s_1^{z^*}} = \mathbf{0}, \boldsymbol{s_2^{z^*}} = \mathbf{0}, \tilde{\boldsymbol{s}}^{\boldsymbol{x}^*} = \mathbf{0}, \tilde{\boldsymbol{s}}^{\boldsymbol{y}^*} = \mathbf{0}, \tilde{\boldsymbol{s}}_{\boldsymbol{1}}^{\boldsymbol{z}^*} = \mathbf{0}$, and $\tilde{\boldsymbol{s}}_{\boldsymbol{2}}^{\boldsymbol{z}^*} = \mathbf{0}$, i.e., no real and integer input excesses, no real and integer output shortfalls, and no real and integer flexible measure slacks in any optimal solution. DMU_o $(\boldsymbol{x}_h, \boldsymbol{y}_h, \boldsymbol{z}_h, \tilde{\boldsymbol{x}}_h, \tilde{\boldsymbol{y}}_h, \tilde{\boldsymbol{z}}_h)$ is inefficient if $\rho^{*mFISBM}_{h} < 1$. This condition means we have the following expression for inefficient DMU_h $(\boldsymbol{x}_h, \boldsymbol{y}_h, \boldsymbol{z}_h, \tilde{\boldsymbol{x}}_h, \tilde{\boldsymbol{y}}_h, \tilde{\boldsymbol{z}}_h)$: $\boldsymbol{x}_h = \boldsymbol{X\lambda}^* + \boldsymbol{s}^{\boldsymbol{x}^*}$, $\boldsymbol{y}_h = \boldsymbol{Y\lambda}^* - \boldsymbol{s}^{\boldsymbol{y}^*}$, $\boldsymbol{z}_h = \boldsymbol{Y\lambda}^* + \boldsymbol{s}_{\boldsymbol{1}}^{\boldsymbol{z}^*} - \boldsymbol{s}_{\boldsymbol{2}}^{\boldsymbol{z}^*}$ where $\boldsymbol{s}_{\boldsymbol{1}}^{\boldsymbol{z}^*}.\boldsymbol{s}_{\boldsymbol{2}}^{\boldsymbol{z}^*} = \mathbf{0}$, $\tilde{\boldsymbol{x}}_h = \tilde{\boldsymbol{X}}\boldsymbol{\lambda}^* + \tilde{\boldsymbol{s}}^{\boldsymbol{x}^*}$, $\tilde{\boldsymbol{y}}_h = \tilde{\boldsymbol{Y}}\boldsymbol{\lambda}^* - \tilde{\boldsymbol{s}}^{\boldsymbol{y}^*}$, $\tilde{\boldsymbol{z}}_h = \tilde{\boldsymbol{Y}}\boldsymbol{\lambda}^* + \tilde{\boldsymbol{s}}_{\boldsymbol{1}}^{\boldsymbol{z}^*} - \tilde{\boldsymbol{s}}_{\boldsymbol{2}}^{\boldsymbol{z}^*}$ where $\tilde{\boldsymbol{s}}_{\boldsymbol{1}}^{\boldsymbol{z}^*}.\tilde{\boldsymbol{s}}_{\boldsymbol{2}}^{\boldsymbol{z}^*} = \mathbf{0}$. Straightforwardly, DMU_h can become efficient by omitting the slacks i.e., $\boldsymbol{x}_h \leftarrow \boldsymbol{x}_h - \boldsymbol{s}^{\boldsymbol{x}^*}, \boldsymbol{y}_h \leftarrow \boldsymbol{y}_h + \boldsymbol{s}^{\boldsymbol{y}^*}, \boldsymbol{z}_h \leftarrow \boldsymbol{z}_h - \boldsymbol{s}_{\boldsymbol{1}}^{\boldsymbol{z}^*} + \boldsymbol{s}_{\boldsymbol{2}}^{\boldsymbol{z}^*} : \boldsymbol{s}_{\boldsymbol{1}}^{\boldsymbol{z}^*} \cdot \boldsymbol{s}_{\boldsymbol{2}}^{\boldsymbol{z}^*} = \mathbf{0}, \tilde{\boldsymbol{x}}_h \leftarrow \tilde{\boldsymbol{x}}_h - \tilde{\boldsymbol{s}}^{\boldsymbol{x}^*}, \tilde{\boldsymbol{y}}_h \leftarrow \tilde{\boldsymbol{y}}_h + \tilde{\boldsymbol{s}}^{\boldsymbol{y}^*}, \tilde{\boldsymbol{z}}_h \leftarrow \tilde{\boldsymbol{z}}_h - \tilde{\boldsymbol{s}}_{\boldsymbol{1}}^{\boldsymbol{z}^*} + \tilde{\boldsymbol{s}}_{\boldsymbol{2}}^{\boldsymbol{z}^*} : \tilde{\boldsymbol{s}}_{\boldsymbol{1}}^{\boldsymbol{z}^*} \cdot \tilde{\boldsymbol{s}}_{\boldsymbol{2}}^{\boldsymbol{z}^*} = \mathbf{0}$. These operations can be called *mFISBM-projections* as the *SBM-projection* in Tone (2001).

The set of DMUs with the corresponding $\boldsymbol{\lambda}^* > \mathbf{0}$ is called reference-set to DMU_h as the SBM DEA model. Furthermore, a DMU is *FISBM-efficient* if and only if it is *mFISBM-efficient* (see Appendix A). Similar to the SBM model (Tone 2001), the formulation of ρ_h^{mFISBM} in Model (7) can be interpreted as the product of input and output inefficiencies or the second term of numerator and denominator, correspondingly. Then, the numerator and denominator evaluate, respectively, the mean reduction rate of inputs and mean expansion rate of outputs considering the optimal role of flexible measures as well. It should be noted that when $\rho^{*mFISBM}_{h} = 1$, the status of the real- and integre-valued flexible measures cannot be declared for DMU_o. As explained by Boďa (2020), this is the case of indefinite and it reports technical efficiency score where no matter what *L*-tuple of $\{0,1\}$ is taken for $\{\boldsymbol{d}, \tilde{\boldsymbol{d}}\}$.

The non-oriented *mFISBM* DEA model can be reformulated as input-oriented (IO) by setting the denominator of the Eq. (7.1) to one and excluding Constraints (7.4), (7.6), (7.10), and (7.12) from the system. In the same way, the output-oriented (OO) *mFISBM* DEA model can be written but, in this case, we maximize the denominator and set the numerator to 1, and remove Constraints (7.5), (7.7), (7.11), and (7.13) from the model. The oriented *mFISBM* technical efficiency scores are optimal values

$\rho^{*}{}_{h}^{mFISBM-IO}$ and $1/\rho^{*}{}_{h}^{mFISBM-OO}$ where $\rho^{*}{}_{h}^{mFISBM-IO} and\ 1/\rho^{*}{}_{h}^{mFISBM-OO} \geq \rho^{*}{}_{h}^{mFISBM}$. Then, there is no need for the *Charnes–Cooper* transformation (explicitly, no need for dealing with multiplying the scalar variable $t > 0$ in slacks) since both objective functions are linear and the optimum solutions are directly reported by the models.

When dealing with large case studies, there may be a concern about the size (total number of decision variables and constraints) of Model (7) compared to Model (5). This could be problematic from the perspective of computational complexity. To handle this issue, we propose Model (8) with less size than Model (7) as follows (hereafter revised *FISBM* or *rFISBM*):

$$[rFISBM]$$

$$\Omega_h^{rFISBM} = Min\ 1 - \frac{\left[\sum_{i\in I^{NI}}\frac{s^{*x}_i}{x_{ih}} + \sum_{i\in I^{I}}\frac{\tilde{s}^{*x}_i}{x_{ih}} + \sum_{k\in K^{NI}}\frac{s^{*z}_{1k}}{z_{kh}} + \sum_{k\in K^{I}}\frac{\tilde{s}^{*z}_{1k}}{z_{kh}} + \sum_{r\in O^{NI}}\frac{s^{*y}_r}{y_{rh}} + \sum_{r\in O^{I}}\frac{\tilde{s}^{*y}_r}{y_{rh}} + \sum_{k\in K^{NI}}\frac{s^{*z}_{2k}}{z_{kh}} + \sum_{k\in K^{I}}\frac{\tilde{s}^{*z}_{2k}}{z_{kh}}\right]}{m+s+p} \quad (8.1)$$

$$s.t. \quad (5.2) - (5.17).$$

Model (8) is also units-invariant and provides an integrated efficiency index (Ω_h^{rFISBM}) ranging from 0 to 1 (see Appendix B). Ω_h^{rFISBM} can be also called monotonically decreasing with respect to input, output, and flexible slacks so that a larger value represents a smaller slack ratio then, better performance in reaching the efficient frontier. However, unlike ρ_h^{mFISBM}, Ω_h^{rFISBM} cannot be construed as the product of input and output inefficiencies. Therefore, the efficiency index calculated by Model (8) cannot be recommended when investigating inefficiency sources is the goal of performance evaluation.

4 Application: The Case of German University Hospitals

In this section, we use a dataset of 28 public university hospitals in Germany[2] in 2017. The data collection was carried out in different research steps including homepages of the hospitals and direct contact (e-mail/telephone inquiries) to the responsible departments and proved to be very cumbersome. For inputs, we consider the number of beds, physicians, and nurses. The number of beds is an integer-valued input measure. However, physicians and nurses are in full-time equivalent (FTE) units, i.e., real values. The number of outpatients and case-mix adjusted discharges for inpatients are designated as integer and real outputs, respectively. However, these two outputs as the major outputs for general hospitals, do not provide teaching function. Therefore, we use the number of medical students as the integer-valued output of the university hospitals. The total number of students enrolled in the university hospital's medical degree programs is reflected in this factor. Since they are not yet trained to practice medicine alongside physicians at a population level, they cannot work in any specialties. This makes

[2] There exist 35 German university hospitals together with their medical faculties. However, due to the lack of availability of data for seven units, they have been excluded from the analysis. A complete list of German university hospitals is available at https://www.uniklinika.de/.

them ineligible to be considered as input (trained staff) for university hospitals. However, the degree to which teaching contributes to the training of highly skilled personnel is also an important component in the academic mission performance of a university hospital. Therefore, the total number of medical students is considered as an output (Ozcan et al. 2010). We also introduce two more flexible measures to represent the teaching function in the efficiency assessment: the number of graduates and third-party funding income. Graduates who have completed their doctorate in medicine (trained) in a university hospital can play the role of either input (an available and qualified resource who can work under the supervision of the faculties or physicians so can affect their productivity) or output (accomplished staff, then a benefit resulting from teaching funding). Third-party funding income can be similarly interpreted in the efficiency evaluation of university hospitals; as input (a form of earnings received) or as output since most research-granting agencies are willing to assign funds to the university hospitals with the supreme impact.

Table 1 represents the data of 28 university hospitals with 3 inputs, 3 outputs, and 2 flexible measures. In the last four rows of the table, the descriptive statistics are reported. The university hospitals considered in this study have on average 1,475 beds which are categorized as the large hospital. They employ more than 25,000 and 34,000 FTE physicians and nurses, respectively. From the output perspective, in total over 2.8 million adjusted inpatient admissions and over 11.4 million outpatient visits occurred. The teaching measures show that about 11 thousand graduates and about 84 thousand medical students in these hospitals where have received over €1.5 billion from the research-granting agencies.

The results of efficiency analysis of the teaching universities obtained from Model (5) (Kordrostami et al. 2019), and the proposed Models (7) and (8) are respectively reported in Tables C1, C2, and C3 in Appendix C. All three models are run under the CRS setting and implemented in *IBM ILOG CPLEX Optimization Studio*. As might be expected, they exhibit differences and share properties in common. University hospitals 2, 6, 8, 15, 21, 23, and 27 are characterized by all three models as efficient DMUs with the optimum slacks of zero. As claimed in Theorem 1, Model (7) will characterize a DMU as efficient if and only if Model (5) characterizes it as efficient. To interpret the integrality, we run the relaxed form of Model (7) in which the integrality is relaxed. Then, we examine the result of an inefficient unit, say university hospital #9. The optimum objective value of the integrality-relaxed Model (7) $\rho^{*}_{9}{}^{relaxed} = 0.7799$ obtained with the intensity optimum weights $\{\lambda_2^* = 0.3189, \lambda_5^* = 0.1753,\ \lambda_{23}^* = 0.0550\}$ and other λ^* are equal to zero. This set of optimum weights results in the reference input (number of beds) $\sum_{j=1}^{28} \lambda_j^* x_{Beds,j} = 1{,}269.0274$ which dominates the integer-valued input target obtained from Model (7), $\tilde{x}_{Beds} = 1280$. Model (7) implies that $\tilde{x}_{Beds} = 1280$ (or 330 units reduction in beds) is a feasible target where is not outside of the real PPS. However, there exist some situations in which the integer-valued reference input is not feasible. For example, consider

university hospital #4. The optimum efficiency score obtained from the integrality-relaxed Model (7), $\rho^{*relaxed}_{4} = 0.8201$. This yields with the intensity optimum weights $\{\lambda^*_6 = 0.7795, \lambda^*_{12} = 0.0245, \lambda^*_{23} = 0.0516\}$ (others are equal to zero). This reports the reference input ("Beds") $\sum_{j=1}^{28} \lambda^*_j x_{Beds,j} = 1{,}173.5570$ which does not dominate the integer-valued input target $\tilde{x}_{Beds} = 1049$. The result is due to the designated status of the flexible measures in the final PPS. For the DMU_4, the real flexible measure "Third-party funding income" is detected as input in the non-integer PPS while it plays the role of output in the integer PPS. Therefore, the PPS may not be comparable in some situations where the flexible measures can play different roles. Slacks of the convex PPS (produced by the non-integer DEA) are usually real-valued amounts and that the optimal integer input/output slacks reported by the models are not constantly a rounding up or down of the real-valued slacks. As reported in Table C2 in Appendix, for example, DMU_4, the integer slacks of "Beds" $\tilde{s}^*_{Beds} = 146$ which is not equal to the rounded up or down of its convex (non-integer) slack $s^*_{Beds} = 121.44$. Incontestably, for the university hospital #19, $\tilde{s}^*_{Beds} = 39$ differs expressively from its corresponding non-integer slack $s^*_{Beds} = 259.58$.

Table 1. Data of 28 German university hospitals in 2017

DMU	***Beds***	***Physicians***	***Nurses***	***Inpatients***	***Outpatients***	***Students***	***Third-party funding income (10^3 €)***	***Graduates***
1	1,517	878.81	1,124.81	79,965.62	245,085	2,339	38,708	330
2	3,011	1,998.50	2,618.66	224,328.58	1,537,233	7,432	153,400	795
3	1,237	725.71	897.95	76,312.64	342,327	2,993	40,524	289
4	1,295	852.98	1,304.83	97,594.78	428,046	2,699	46,882	315
5	1,303	793.38	1,035.13	89,673.48	377,545	3,232	31,678	386
6	1,378	823.63	1,353.20	113,658.93	517,851	3,212	37,249	472
7	1,260	831.50	1,035.27	85,042.01	226,331	1,885	35,505	240
8	1,297	709.43	936.18	81,876.35	276,610	3,416	39,286	562
9	1,610	1,067.65	1,274.21	95,335.71	578,049	3,090	76,200	459
10	1,554	845.50	1,294.47	89,431.96	214,921	2,861	52,169	346
11	919	436.70	715.07	56,000.14	17,095	1,598	21,248	204
12	984	521.20	762.04	58,401.14	169,302	2,110	11,473	199
13	1,436	1,118.10	1,468.10	96,848.74	337,455	3,347	79,946	405
14	1,520	834.93	1,314.90	118,360.73	459,719	2,581	91,368	328
15	1,988	1,471.50	1,573.59	137,557.42	1,093,862	3,398	105,465	478
16	1,396	741.83	1,130.74	93,933.26	463,361	2,334	27,000	315
17	1,464	790.14	1,210.11	101,525.98	329,189	3,338	98,513	402
18	1,345	773.03	1,042.71	86,053.41	296,937	2,758	42,977	381
19	1,662	978.73	1,314.89	107,254.52	269,380	3,417	45,800	442
20	1,352	621.37	757.41	69,898.95	214,535	1,483	41,913	211
21	2,050	1,182.41	1,856.37	155,754.00	834,985	5,616	96,770	676
22	1,091	949.94	1,071.65	86,508.14	254,462	1,715	45,585	496
23	1,457	948.44	1,228.14	146,479.82	391,521	2,777	47,620	293
24	833	626.85	898.83	69,979.62	154,657	2,010	22,220	268
25	2,196	1,156.13	1,996.66	160,488.88	302,263	3,566	63,900	452
26	1,559	850.40	1,065.25	98,409.18	383,947	2,998	90,400	541
27	1,150	741.23	813.20	79,745.76	247,370	2,736	55,200	340
28	1,438	862.50	1,294.44	98,027.93	489,027	2,842	36,388	366
Sum	41,302.0	25,132.5	34,388.8	2,854,447.7	11,453,065.0	83,783.0	1,575,387.0	10,991.0
Average	1,475.1	897.6	1,228.2	101,944.6	409,038.0	2,992.3	56,263.8	392.5
StD	430.6	301.6	408.3	35,664.6	305,287.9	1,185.6	31,836.8	138.7
Min	833.0	436.7	715.1	56,000.1	17,095.0	1,483.0	11,473.0	199.0
Max	3,011.0	1,998.5	2,618.7	224,328.6	1,537,233.0	7,432.0	153,400.0	795.0

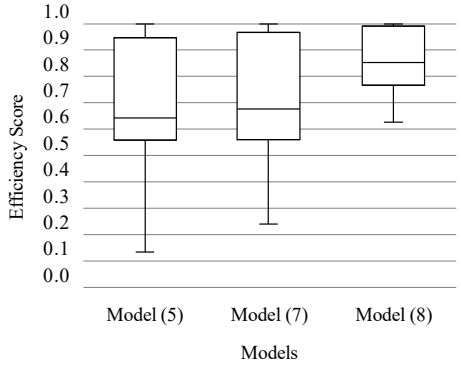


Figure 2. Efficiency scores calculated by Models (5), (7), and (8)

In Tables C1, C2, and C3 in Appendix C, d_k^* and $\tilde{d}_k^*$ indicate the roles of "Third-party funding income" and "Graduates" in the final PPS, respectively. In Model (5), 19 out of the 28 university hospitals treat these two flexible measures as output i.e., the majority treats both as output. The same results are reported by Model (7) where 18 and 20 DMUs determine the status of both "Third-party funding income" and "Graduates" as output, correspondingly. However, Model (8) assigns different optimal designations for these two flexible measures so that only 9 and 7 university hospitals identify the role of "Third-party funding income" and "Graduates" respectively as output, i.e., the majority of 19 and 21 DMUs treat them as input. This can explain the difference between the efficiency scores calculated by Models (5) and (7) with those calculated by Model (8) as illustrated in Figure 2. The inefficiency scores obtained from all three models are asymmetrically distributed since the medians of inefficiency scores are not in the middle of the boxes, and the whiskers are not about the same on the upper and lower sides. These boxes are also advantageous for offering a visual indicator of the variability of inefficiencies. The minimum of 0.6263, the first quartile at 0.7660, and a standard deviation equal to 0.1095, all signify the limited discriminative power of Model (8)'s inefficiencies in this case. However, the situation changes with Models (5) and (7) where the longer boxes show more dispersed and scattered inefficiency scores. Since the median lines of Model (5) ($= 0.6434$) and Model (7) ($= 0.6760$) are close to each other, there is likely to be no difference between the efficiency scores of these two models. On the other hand, the median line of Model (5) that sits above 0.8534, represents the possibility of a difference between the inefficiency scores calculated by this model and the others. None inefficiency score is detected as the outlier. In other words, the lowest inefficiency score computed by the models is within one and a half interquartile range of its 25th-percentile, and the maximum efficiency score (1.0) is within one and a half of its 75th-percentile.

To analyze the magnitudes and sources of inefficiency regarding the corresponding inputs/outputs for each inefficient university hospital, the inefficiency scores can be decomposed using the optimal solution obtained from the models as exhibited in Table 2. This decomposition provides managers or policy-makers with enlightening information about how to become an efficient DMU by examining the magnitudes and sources of inefficiency. As might be expected, the majority of the

inefficiency sources identified by Models (7) and (8) are input inefficiency since we run the input-oriented form of Model (7) and Model (8) also calculates the scores by minimizing the Ω_o^{rFISBM} (Eq. (8.1)). From Table 2, we can see 17 out of 21 inefficiencies are caused by input inefficiencies in both Models (7) and (8). However, this is different for Model (5) where 18 out of 21 inefficiencies are attributed to the output inefficiency. Part of the clarification for the distinct results may be that Model (5) is additive and its objective function maximizes the summation of slacks (see Eq. (5.1)) instead of targeting input/output inefficiencies. Those four university hospitals (namely, 5, 12, 16, and 24) in which the output inefficiency is identified as the main source of inefficiency share one significant property in common. They all have a considerable amount of slacks of the flexible measure "Third-party funding income" that is designated as output ($d_k^* = 1$) in the optimum solutions obtained from all three models (see Tables C1, C2, and C3 in Appendix C). This indicates the significant shortage in the third-party funding income dominates other inefficiencies.

Table 2. Inefficiency decomposition

DMU	*Model (5)*			*Model (7)*			*Model (8)*		
	Input Ineff	Output Ineff	Dominant	Input Ineff	Output Ineff	Dominant	Input Ineff	Output Ineff	Dominant
1	0.0000	0.6605	Output Ineff	0.3092	0.0000	Input Ineff	0.2957	0.0000	Input Ineff
2	-	-	-	-	-	-	-	-	-
3	0.0292	0.1511	Output Ineff	0.0700	0.0122	Input Ineff	0.0700	0.0122	Input Ineff
4	0.0752	0.2428	Output Ineff	0.1673	0.0021	Input Ineff	0.1799	0.0000	Input Ineff
5	0.0265	0.2363	Output Ineff	0.0645	0.1207	Output Ineff	0.0645	0.1207	Output Ineff
6	-	-	-	-	-	-	-	-	-
7	0.0129	0.6903	Output Ineff	0.2678	0.0000	Input Ineff	0.2494	0.0000	Input Ineff
8	-	-	-	-	-	-	-	-	-
9	0.2034	0.0318	Input Ineff	0.2183	0.0000	Input Ineff	0.2200	0.0000	Input Ineff
10	0.0249	0.6076	Output Ineff	0.2410	0.0000	Input Ineff	0.2472	0.0000	Input Ineff
11	0.0881	3.6791	Output Ineff	0.1835	0.0348	Input Ineff	0.1663	0.0000	Input Ineff
12	0.0074	1.0592	Output Ineff	0.1505	0.4298	Output Ineff	0.1505	0.4298	Output Ineff
13	0.1966	0.2590	Output Ineff	0.2813	0.0000	Input Ineff	0.2386	0.0000	Input Ineff
14	0.1704	0.1366	Input Ineff	0.1853	0.0000	Input Ineff	0.1866	0.0549	Input Ineff
15	-	-	-	-	-	-	-	-	-
16	0.0203	0.4655	Output Ineff	0.1334	0.2007	Output Ineff	0.1334	0.2007	Output Ineff
17	0.1082	0.2251	Output Ineff	0.1644	0.0000	Input Ineff	0.1887	0.0442	Input Ineff
18	0.0884	0.3586	Output Ineff	0.1705	0.0000	Input Ineff	0.1705	0.0000	Input Ineff
19	0.0636	0.6312	Output Ineff	0.1731	0.0000	Input Ineff	0.1731	0.0000	Input Ineff
20	0.1411	0.4423	Output Ineff	0.3012	0.0000	Input Ineff	0.3067	0.0000	Input Ineff
21	-	-	-	-	-	-	-	-	-
22	0.2009	0.4104	Output Ineff	0.3737	0.0000	Input Ineff	0.3737	0.0000	Input Ineff
23	-	-	-	-	-	-	-	-	-
24	0.1272	0.4113	Output Ineff	0.1335	0.1383	Output Ineff	0.1335	0.1383	Output Ineff
25	0.0480	0.4826	Output Ineff	0.1682	0.0000	Input Ineff	0.1816	0.0000	Input Ineff
26	0.2039	0.1230	Input Ineff	0.2469	0.0000	Input Ineff	0.2469	0.0000	Input Ineff
27	-	-	-	-	-	-	-	-	-
28	0.0128	0.3692	Output Ineff	0.1577	0.1149	Input Ineff	0.1577	0.1149	Input Ineff

Ineff: Inefficiency

Now turning to the teaching function, we can see from the reported slacks in Tables C2 and C3 that "Third-party funding income" as one of the teaching proxies have the maximum ratio of slacks (either input excesses or output shortfalls) in all university hospitals except DMU_7 in Model (7) where excesses in inputs ("Beds", "Physicians", and "Nurses") dominate other inefficiencies. This specifies in almost all the evaluating university hospitals in Germany, teaching inefficiency dominates the general inefficiency. As is now apparent the same result is not seen in the optimum solutions calculated by Model (5) in Table C1 in Appendix C. In this model, the slacks of "Third-party funding income" are as

well substantial while shortages in "Outpatients" are identified as the dominator. This disparity can be due to the fact that Model (5) is additive and its objective function cannot be explained as the inefficiency ratio.

It is not easy to evaluate our findings in the light of other studies since there are no recent and comparable studies on the university hospital performance assessment, especially in Germany. However, studies dealing with the efficiency of university hospitals (as discussed in Section 2) have already pointed out differences in efficiency between teaching and non-teaching hospitals. Generally speaking, university hospitals are not able to compete with non-teaching counterparts since they pursue different goals.

5 Conclusions

In this study, we advance the SBM DEA model proposed by Tone (2001) to consider the real circumstances of the integer nature of certain measures whose status can be flexibly designated. Besides, we develop a revision to the additive model developed by Kordrostami et al. (2019) to make the model report of a non-negative inefficiency index with an upper limit of one. Then, the optimal solutions derived from the proposed and revised models are investigated in comparison with Kordrostami's solutions. This is illustrated by the performance analysis of 28 university hospitals in Germany. In this case study, in addition to the patient care function, the teaching function of the units is captured in the PPS by introducing two flexible measures containing one real-valued ("Third-party funding income") and one integer-valued ("Graduates") as well as one integer-valued output ("Students"). In this application, the inclusion of the integrality constraints leads to more valid slacks, i.e., ensures to lie within the integer PPS and to not be dominated by any other feasible units. The proposed model describes more reliable and discriminated inefficiency scores from which a more successful ordering of the university hospitals can be originated.

From a practical viewpoint, the decomposition of inefficiencies provides hospital managers, local and national health authorities some informative insights on the source and magnitude of the inefficiency of German university hospitals. The significant shortage in the third-party funding that university hospitals receive as a form of revenue is identified as the main source of inefficiency. Having this fact in mind that most research-granting organizations (e.g., German Research Foundation) consider the university hospitals with the greatest impact, it can be concluded that targeting research missions might boost the efficiency of German university hospitals. A reconsideration might therefore be required in the university hospital performance management. The enormous public funds that flow into medical education should be allocated more according to efficiency aspects. Now that health care is under increasing pressure to be more efficient due to the introduction of a more results-oriented reimbursement system, similar instruments should also be used for the reimbursement of the academic

mission. The proposed SBM DEA model could be used as an accompanying controlling and monitoring instrument. At the same time, in order to avoid cross-subsidies between academic and patient care missions in university hospitals, more transparency is urgently needed by applying a performance assessment approach that allows both missions to be efficiently combined under one roof. Since high-quality teaching cannot be separated from patient care, this realization can give politicians a clear mandate to find a solution to this dilemma. The proposed model could be a suitable monitoring approach for this path, taking into account further comparative parameters and the necessary modifications in the dataset used in the analysis such as identifying new measures.

A weakness of the conceptualized model is the lack of the quality of patient care in the analysis. However, these datasets are usually classified and are not publicly available. In addition, an attempt should be made to integrate the other university hospitals into the investigation and to conduct an analysis over a longer period of time. A longitudinal study would allow statements on the development of efficiency of individual university hospitals, for instance, in order to assess the efficiency effect of mergers. As a real example, the German Federal Cartel Office[3] has recently explicated plans to merge the cardiological and cardiosurgical services of the *Charité* and *Deutsches Herzzentrum Berlin* and establish the heart center *Deutsches Herzzentrum der Charité* (Bundeskartellamt 2021). Furthermore, from a theoretical perspective, one of the limitations of this study that can be addressed in the future may be extending the present model by incorporating the perspective of the radial characteristics of measure in inefficiency sources. This leads to bring the effects of inputs/outputs that are subject to change proportionally.

References

Amirteimoori, A., & Emrouznejad, A. (2012). Notes on "Classifying inputs and outputs in data envelopment analysis". *Applied Mathematics Letters, 25*(11): 1625–1628.

Amirteimoori, A., Emrouznejad, A., & Khoshandam, L. (2013). Classifying flexible measures in data envelopment analysis: A slack-based measure. *Measurement, 46*(10): 4100–4107.

Andersen, P., & Petersen, N. Christian (1993). A Procedure for Ranking Efficient Units in Data Envelopment Analysis. *Management Science, 39*(10): 1261–1264.

Arana-Jiménez, M., Sánchez-Gil, M., Younesi, A., & Lozano, S. (2020). Integer interval DEA: An axiomatic derivation of the technology and an additive, slacks-based model. *Fuzzy Sets and Systems*.

Beasley, J. E. (1990). Comparing university departments. *Omega, 18*(2): 171–183.

Boďa, M. (2020). Classifying flexible measures in data envelopment analysis: A slacks-based measure – A comment. *Measurement, 150*: p. 107045.

Bundeskartellamt (2021). Bundeskartellamt clears merger between Charité and Deutsches Herzzentrum Berlin. Bundeskartellamt/Hospital Sector. Bonn. Available online at https://www.bundeskartellamt.de/SharedDocs/Publikation/EN/Pressemitteilungen/2021/07_0

[3] In German: *Bundeskartellamt*

6_2021_Charite_DHZ.pdf?__blob=publicationFile&v=4, updated on 7/6/2021, checked on 8/12/2021.

Charnes, A., Cooper, W. W., & Rhodes, E. (1978). Measuring the efficiency of decision making units. *European Journal of Operational Research, 2*(6): 429–444.

Cook, W. D., Green, R. H., & Zhu, J. (2006). Dual-role factors in data envelopment analysis. *IIE Transactions, 38*(2): 105–115.

Cook, W. D., & Zhu, J. (2007). Classifying inputs and outputs in data envelopment analysis. *European Journal of Operational Research, 180*(2): 692–699.

Du, J., Chen, C.-M., Chen, Y., Cook, W. D., & Zhu, J. (2012). Additive super-efficiency in integer-valued data envelopment analysis. *European Journal of Operational Research, 218*(1): 186–192.

Ghiyasi, M., & Cook, W. D. (2021). Classifying dual role variables in DEA: The case of VRS. *Journal of the Operational Research Society, 72*(5): 1183–1190.

Grosskopf, S., Margaritis, D., & Valdmanis, V. (2001). Comparing Teaching and Non-teaching Hospitals: A Frontier Approach (Teaching vs. Non-teaching Hospitals). *Health Care Management Science, 4*(2): 83–90.

Grosskopf, S., Margaritis, D., & Valdmanis, V. (2004). Competitive effects on teaching hospitals. *European Journal of Operational Research, 154*(2): 515–525.

Jie, T., Yan, Q., & Xu, W. (2015). A technical note on "A note on integer-valued radial model in DEA". *Computers & Industrial Engineering, 87*: 308–310.

Kazemi Matin, R., & Kuosmanen, T. (2009). Theory of integer-valued data envelopment analysis under alternative returns to scale axioms. *Omega, 37*(5): 988–995.

Khezrimotlagh, D., Salleh, S., & Mohsenpour, Z. (2013a). A new robust mixed integer-valued model in DEA. *Applied Mathematical Modelling, 37*(24): 9885–9897.

Khezrimotlagh, D., Salleh, S., & Mohsenpour, Z. (2013b). A note on integer-valued radial model in DEA. *Computers & Industrial Engineering, 66*(1): 199–200.

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.

Kordrostami, S., Amirteimoori, A., & Jahani Sayyad Noveiri, M. (2019). Inputs and outputs classification in integer-valued data envelopment analysis. *Measurement, 139*: 317–325.

Kourtzidis, S., Matousek, R., & Tzeremes, N. G. (2021). Modelling a multi-period production process: Evidence from the Japanese regional banks. *European Journal of Operational Research, 294*(1): 327–339.

Kuosmanen, T., Keshvari, A., & Matin, R. Kazemi (2015). Discrete and Integer Valued Inputs and Outputs in Data Envelopment Analysis. In Joe Zhu (Ed.): Data Envelopment Analysis: A Handbook of Models and Methods. Boston, MA: Springer US: 67–103.

Kuosmanen, T., & Matin, R. Kazemi (2009). Theory of integer-valued data envelopment analysis. *European Journal of Operational Research, 192*(2): 658–667.

Lobo, M. Stella Castro, Ozcan, Y. A., Lins, M. P. Estellita, Silva, A. Cristina M., & Fiszman, R. (2014). Teaching hospitals in Brazil: Findings on determinants for efficiency. *International Journal of Healthcare Management, 7*(1): 60–68.

Lozano, S., & Villa, G. (2006). Data envelopment analysis of integer-valued inputs and outputs. *Computers & Operations Research, 33*(10): 3004–3014.

Lozano, S., & Villa, G. (2007). Integer Dea Models. In Joe Zhu, Wade D. Cook (Eds.): Modeling Data Irregularities and Structural Complexities in Data Envelopment Analysis. Boston, MA: Springer US: 271–289.

Ozcan, Y. A., Lins, M. E., Lobo, M. Stella C., da Silva, A. Cristina M., Fiszman, R., & Pereira, B. B. (2010). Evaluating the performance of Brazilian university hospitals. *Annals of Operations Research, 178*(1): 247–261.

Schneider, A. Maren, Oppel, E.-M., & Schreyögg, J. (2020). Investigating the link between medical urgency and hospital efficiency - Insights from the German hospital market. *Health Care Management Science, 23*(4): 649–660.

Sedighi Hassan Kiyadeh, M., Saati, S., & Kordrostami, S. (2019). Improvement of models for determination of flexible factor type in data envelopment analysis. *Measurement, 137*: 49–57.

Stefaniec, A., Hosseini, K., Xie, J., & Li, Y. (2020). Sustainability assessment of inland transportation in China: A triple bottom line-based network DEA approach. *Transportation Research Part D: Transport and Environment, 80*: p. 102258.

Tohidi, G., & Matroud, F. (2017). A new non-oriented model for classifying flexible measures in DEA. *Journal of the Operational Research Society, 68*(9): 1019–1029.

Toloo, M. (2009). On classifying inputs and outputs in DEA: A revised model. *European Journal of Operational Research, 198*(1): 358–360.

Toloo, M. (2012). Alternative solutions for classifying inputs and outputs in data envelopment analysis. *Computers & Mathematics with Applications, 63*(6): 1104–1110.

Toloo, M., Ebrahimi, B., & Amin, G. R. (2021). New data envelopment analysis models for classifying flexible measures: The role of non-Archimedean epsilon. *European Journal of Operational Research, 292*(3): 1037–1050.

Tone, K. (2001). A slacks-based measure of efficiency in data envelopment analysis. *European Journal of Operational Research, 130*(3): 498–509.

Tone, K. (2017). Advances in DEA Theory and Applications. Chichester, UK: John Wiley & Sons, Ltd.

Tulkens, H. (2006). On FDH Efficiency Analysis: Some Methodological Issues and Applications to Retail Banking, Courts and Urban Transit. In Parkash Chander, Jacques Drèze, C. Knox Lovell, Jack Mintz (Eds.): Public goods, environmental externalities and fiscal competition. Boston, MA: Springer US: 311–342.

Villano, R. A., & Tran, C.-D. T. T. (2018). Performance of private higher education institutions in Vietnam: evidence using DEA-based bootstrap directional distance approach with quasi-fixed inputs. *Applied Economics, 50*(55): 5966–5978.

Appendix A.

Theorem 1. A DMU is *FISBM-efficient* if and only if it is *mFISBM-efficient*, i.e., $\tau_h^{*FISBM} = 0 \leftrightarrow \rho_h^{*mFISBM} = 1$.

Proof. $\tau_h^{*FISBM} = 0$ if and only if the optimum value of all inputs, outputs, and flexible slacks be equal to 0 considering the nonnegativity condition imposed by Constraint (5.16), i.e., $\boldsymbol{s}^* = (\boldsymbol{s}^{*x}, \boldsymbol{s}^{*y}, \tilde{\boldsymbol{s}}^{*x}, \tilde{\boldsymbol{s}}^{*y}, \boldsymbol{s}^{*z}_1, \boldsymbol{s}^{*z}_2) = \mathbf{0}$. By replacing this solution into Model (7), we have $\boldsymbol{\delta}^x \leq \mathbf{0}$ and $\boldsymbol{\delta}^x \geq \mathbf{0}$ from Constraint (7.8) which results in $\boldsymbol{\delta}^x = \mathbf{0}$**.** The same results about $\boldsymbol{\delta}^y, \boldsymbol{\delta}^z$ and $\tilde{\boldsymbol{\delta}}^x, \tilde{\boldsymbol{\delta}}^y, \tilde{\boldsymbol{\delta}}^z$ would be achieved from Constraints (7.4), (7.5), (7.6), (7.10), (7.11), (7.12), and (7.13), respectively. On the

other hand, we know that DMU_h in *mFISBM* model is called efficient if and only if $\rho^{*}{}_{h}^{mFISBM} = 1$. This stipulation is equal to $\boldsymbol{\delta}^{*} = \left(\boldsymbol{\delta}^{*x}, \boldsymbol{\delta}^{*y}, \boldsymbol{\delta}^{*z}, \tilde{\boldsymbol{\delta}}^{*x}, \tilde{\boldsymbol{\delta}}^{*y}, \tilde{\boldsymbol{\delta}}^{*z}\right) = \mathbf{0}$. This completes the proof. ■

Appendix B.

Theorem 2. $\Omega_o^{rFISBM} \leq 1$.

Proof. Assume a solution of DMU_h in which all input excesses (including the flexible measures designated as input) are equal to the corresponding utilized inputs (or their maximum values) i.e., $\left\{\boldsymbol{s}^{*x} = \boldsymbol{x}_h, \tilde{\boldsymbol{s}}^{*x} = \tilde{\boldsymbol{x}}_h, \boldsymbol{s}_1^{*z} = \boldsymbol{w}_h, \tilde{\boldsymbol{s}}_1{}^{*z} = \tilde{\boldsymbol{w}}_h\right\}$. This follows that $0 \leq (m+p)^{-1}\left[\sum_{i\in I^{NI}} \frac{s^{*x}_i}{x_{ih}} + \sum_{i\in I^{I}} \frac{\tilde{s}^{*x}_i}{x_{ih}} + \sum_{k\in K^{NI}} \frac{s^{*z}_{1k}}{z_{kh}} + \sum_{k\in K^{I}} \frac{\tilde{s}^{*z}_{1k}}{z_{kh}}\right] \leq 1$. However, in the case of outputs, the maximum values for output shortfalls cannot be defined since any output slacks can exceed the corresponding produced outputs $\left\{\mathbf{0} \leq \boldsymbol{s}^{*y}, \mathbf{0} \leq \tilde{\boldsymbol{s}}^{*y}, \mathbf{0} \leq \boldsymbol{s}_2^{*z}, \mathbf{0} \leq \tilde{\boldsymbol{s}}_2{}^{*z}\right\}$, however, it always holds $0 \leq (s+p)^{-1}\left[\sum_{r\in O^{NI}} \frac{s^{*y}_r}{y_{rh}} + \sum_{r\in O^{I}} \frac{\tilde{s}^{*y}_r}{y_{rh}} + \sum_{k\in K^{NI}} \frac{s^{*z}_{2k}}{z_{kh}} + \sum_{k\in K^{I}} \frac{\tilde{s}^{*z}_{2k}}{z_{kh}}\right]$. Since the inefficiency scores are non-negative ($0 \leq \Omega_h^{rFISBM}$), this limits the upper bound of the summation of output mix inefficiencies so that the ratio of average input and output inefficiencies cannot take more than 1. It can reach the upper limit, $\Omega^{*}{}_{h}^{rFISBM} = 1$, only if slacks are equal to their minimum values defined by Eq. (5.16), i.e., $\{\boldsymbol{s}^{*x} = \mathbf{0}, \boldsymbol{s}^{*y} = \mathbf{0}, \tilde{\boldsymbol{s}}^{*x} = \mathbf{0}, \tilde{\boldsymbol{s}}^{*y} = \mathbf{0}, \boldsymbol{s}^{*z} = \mathbf{0}, \tilde{\boldsymbol{s}}^{*z} = \mathbf{0}\}$ which is also a feasible solution for Model (8). ■

Appendix C.

Table C1. Results of efficiency analysis of university hospitals via Model (5)

DMU	τ_h^{*FISBM}	***Real Slacks***						***Integer Slacks***						
		Physicians	*Nurses*	*Inpatients*	*Third-party funding income as input*	*Third-party funding income as output*	d_k^*	*Beds*	*Outpatients*	*Students*	*Graduates as input*	*Graduates as output*	$\tilde{d}_k^*$	ζ_h^{*FISBM}
1	3.303	0.00	0.00	21,132.69	0.00	45,754.35	1	0	247,811	843	0	160	1	0.4760
2	0.000	0.00	0.00	0.00	0.00	0.00	1	0	0	0	0	0	0	1.0000
3	0.843	39.52	0.00	2,310.42	0.00	5,683.70	1	41	0	0	0	169	1	0.7849
4	1.440	105.99	132.17	794.97	0.00	14,250.33	1	0	99,474	848	0	112	1	0.6798
5	1.261	63.19	0.00	214.45	0.00	28,063.17	1	0	27,475	0	0	85	1	0.7061
6	0.000	0.00	0.00	0.00	0.00	0.00	0	0	0	0	0	0	0	1.0000
7	3.490	32.22	0.00	5,129.48	0.00	30,963.51	1	0	332,770	1,042	0	119	1	0.4614
8	0.000	0.00	0.00	0.00	0.00	0.00	1	0	0	0	0	0	1	1.0000
9	1.112	230.31	176.43	0.00	12,816.52	0.00	0	347	54,914	1	128	0	0	0.6487
10	3.113	0.00	0.00	19,489.45	0.00	16,459.12	1	116	389,319	1,024	0	116	1	0.4894
11	18.660	0.00	49.78	0.00	0.00	14,094.23	1	179	296,223	395	0	32	1	0.1328
12	5.318	0.00	0.57	6,885.81	0.00	51,051.96	1	21	52,676	0	0	83	1	0.3600
13	1.823	259.89	58.42	21,558.21	41,107.25	0.00	0	0	201,961	1	0	87	1	0.6381
14	1.228	0.00	9.65	0.00	52,655.75	0.00	0	149	23,974	489	0	100	1	0.7299
15	0.000	0.00	0.00	0.00	0.00	0.00	0	0	0	0	0	0	1	1.0000
16	2.388	0.00	0.13	1,900.89	0.00	41,436.56	1	85	0	1,010	0	107	1	0.5563
17	1.333	0.00	0.00	572.29	37,050.27	0.00	0	83	190,826	421	0	76	1	0.7279
18	1.788	0.00	29.46	744.06	0.00	16,366.99	1	180	297,582	118	73	0	0	0.6710
19	2.779	0.00	32.12	2,631.27	0.00	29,331.48	1	187	483,372	224	52	0	0	0.5740
20	2.333	45.75	0.00	0.00	10,492.57	0.00	0	325	11,989	1,121	0	202	1	0.5955
21	0.000	0.00	0.00	0.00	0.00	0.00	1	0	0	0	0	0	1	1.0000
22	2.445	228.37	128.18	0.00	0.00	6,331.82	1	0	253,996	865	220	0	0	0.5666
23	0.000	0.00	0.00	0.00	0.00	0.00	1	0	0	0	0	0	0	1.0000
24	2.154	113.40	167.11	0.00	0.00	13,546.93	1	0	160,147	0	38	0	0	0.6185
25	2.557	0.00	177.77	0.00	0.00	53,178.41	1	121	387,149	572	0	63	1	0.5383
26	1.388	26.38	0.00	0.00	35,009.23	0.00	0	280	141,659	0	228	0	0	0.6149
27	0.000	0.00	0.00	0.00	0.00	0.00	0	0	0	0	0	0	1	1.0000
28	1.885	33.12	0.10	10,781.51	0.00	31,896.10	1	0	92,511	1,067	0	108	1	0.6143

Table C2. Results of efficiency analysis of university hospitals via input-oriented Model (7)

DMU	$\rho_h^{*mFISBM-IO}$	***Real Slacks***					***Integer Slacks***						
		Physicians	*Nurses*	*Inpatients*	*Third-party funding income as input*	*Third-party funding income as output*	d_k^*	*Beds*	*Outpatients*	*Students*	*Graduates as input*	*Graduates as output*	$\tilde{d}_k^*$
1	0.3730	278.03	320.53	0.00	0.00	0.00	1	495	0	0	0	0	1
2	1.0000	0.00	0.00	0.00	0.00	0.00	1	0	0	0	0	0	1
3	0.8505	87.93	0.00	817.04	0.00	2,034.74	1	110	0	0	0	0	1
4	0.6927	141.97	290.51	0.00	0.00	486.24	1	146	0	0	0	0	1
5	0.8687	98.34	0.46	0.00	0.00	19,115.53	1	90	0	0	0	0	1
6	1.0000	0.00	0.00	0.00	0.00	0.00	1	0	0	0	0	0	1
7	0.4753	242.54	275.07	0.00	0.00	0.00	1	310	0	0	0	0	1
8	1.0000	0.00	0.00	0.00	0.00	0.00	1	0	0	0	0	0	1
9	0.5533	236.00	194.26	0.00	16,860.89	0.00	0	330	0	0	134	0	0
10	0.5021	183.52	301.66	0.00	0.00	0.00	1	424	0	0	0	0	1
11	0.5919	31.62	152.13	0.00	0.00	3,699.12	1	244	0	0	0	0	1
12	0.6660	76.90	66.68	0.00	0.00	24,656.44	1	213	0	0	0	0	1
13	0.5785	353.75	434.63	0.00	35,435.49	0.00	0	100	0	0	0	0	1
14	0.6674	0.23	114.65	0.00	47,997.54	0.00	0	195	0	0	0	0	1
15	1.0000	0.00	0.00	0.00	0.00	0.00	1	0	0	0	0	0	1
16	0.7331	2.04	183.74	0.00	0.00	21,677.36	1	284	0	0	52	0	0
17	0.6975	38.12	72.32	0.00	41,829.17	0.00	0	183	0	0	0	0	1
18	0.6648	57.59	73.33	0.00	17,251.91	0.00	0	126	0	0	81	0	0
19	0.6646	86.88	39.81	0.00	21,297.25	0.00	0	39	0	0	114	0	0
20	0.3144	139.27	98.86	0.00	17,769.25	0.00	0	576	0	0	0	0	1
21	1.0000	0.00	0.00	0.00	0.00	0.00	1	0	0	0	0	0	1
22	0.2394	376.22	327.58	0.00	15,517.20	0.00	0	211	0	0	314	0	0
23	1.0000	0.00	0.00	0.00	0.00	0.00	1	0	0	0	0	0	1
24	0.7329	128.27	171.26	0.00	0.00	12,295.51	1	7	0	0	35	0	0
25	0.6339	70.03	436.79	0.00	11,565.67	0.00	0	467	0	0	0	0	1
26	0.4978	76.81	24.63	0.00	52,067.65	0.00	0	299	0	0	191	0	0
27	1.0000	0.00	0.00	0.00	0.00	0.00	1	0	0	0	0	0	1
28	0.6846	107.60	255.90	0.00	0.00	16,717.84	1	247	0	0	50	0	0

Table C3. Results of efficiency analysis of university hospitals via Model (8)

DMU	$\Omega_h^{*rFISBM}$	***Real Slacks***					***Integer Slacks***						
		Physicians	*Nurses*	*Inpatients*	*Third-party funding income as input*	*Third-party funding income as output*	d_k^*	*Beds*	*Outpatients*	*Students*	*Graduates as input*	*Graduates as output*	$\tilde{d}_k^*$
1	0.7043	241.79	233.92	0.00	16,521.33	0.00	0	403	0	0	100	0	0
2	1.0000	0.00	0.00	0.00	0.00	0.00	0	0	0	0	0	0	0
3	0.9580	87.93	0.00	817.04	0.00	2,034.74	1	110	0	0	0	0	1
4	0.8561	149.01	168.98	0.00	15,054.90	0.00	0	122	0	0	0	0	1
5	0.9613	98.34	0.46	0.00	0.00	19,115.53	1	90	0	0	0	0	1
6	1.0000	0.00	0.00	0.00	0.00	0.00	0	0	0	0	0	0	0
7	0.7506	241.97	248.51	0.00	9,865.38	0.00	0	305	0	0	47	0	0
8	1.0000	0.00	0.00	0.00	0.00	0.00	1	0	0	0	0	0	0
9	0.7800	239.28	190.01	0.00	18,973.93	0.00	0	341	0	0	122	0	0
10	0.7528	100.17	228.10	0.00	31,822.90	0.00	0	196	0	0	71	0	0
11	0.8337	0.00	101.41	0.00	6,322.95	0.00	0	153	0	0	46	0	0
12	0.9097	76.90	66.68	0.00	0.00	24,656.44	1	213	0	0	0	0	1
13	0.7614	279.35	333.13	0.00	50,563.19	0.00	0	0	0	0	34	0	0
14	0.8507	9.80	54.59	0.00	52,680.10	0.00	0	177	0	0	0	72	1
15	1.0000	0.00	0.00	0.00	0.00	0.00	0	0	0	0	0	0	0
16	0.8932	2.04	183.74	0.00	0.00	21,677.36	1	284	0	0	52	0	0
17	0.8490	0.22	161.30	0.00	54,756.14	0.00	0	96	0	0	0	71	1
18	0.8295	57.59	73.33	0.00	17,251.91	0.00	0	126	0	0	81	0	0
19	0.8269	86.88	39.81	0.00	21,297.25	0.00	0	39	0	0	114	0	0
20	0.6933	143.81	133.31	0.00	17,390.89	0.00	0	609	0	0	55	0	0
21	1.0000	0.00	0.00	0.00	0.00	0.00	1	0	0	0	0	0	1
22	0.6263	375.95	327.28	0.00	15,587.54	0.00	0	210	0	0	314	0	0
23	1.0000	0.00	0.00	0.00	0.00	0.00	1	0	0	0	0	0	0
24	0.8932	128.27	171.26	0.00	0.00	12,295.51	1	7	0	0	35	0	0
25	0.8184	42.31	508.89	0.00	15,647.05	0.00	0	389	0	0	88	0	0
26	0.7531	76.81	24.63	0.00	52,067.65	0.00	0	299	0	0	191	0	0
27	1.0000	0.00	0.00	0.00	0.00	0.00	0	0	0	0	0	0	0
28	0.8738	107.60	255.90	0.00	0.00	16,717.84	1	247	0	0	50	0	0

Appendix IV. Analyzing the Relative Efficiency of Internationalization in the University Business Model: The Case of Germany

Jonah M. Otto[a], Mansour Zarrin[b], Dominik Wilhelm[a] and Jens O. Brunner[b]

[a] *Chair of Management and Organization, Faculty of Business and Economics, University of Augsburg, Augsburg, Germany*

[b] *Chair of Healthcare Operations/Health Information Management – UNIKA-T, Center for International Relations, Faculty of Business and Economics, University of Augsburg, Augsburg, Germany*

The printed version is a pre-print of an article published in *Studies in Higher Education*. The final authenticated version is available online at: https://doi.org/10.1080/03075079.2021.1896801.

Status: Published in Studies in Higher Education, Not categorized.

Otto, J. M., Zarrin, M., Wilhelm, D., & Brunner, J. O. (2021). Analyzing the relative efficiency of internationalization in the university business model: the case of Germany. *Studies in Higher Education*, *46*(5), 938-950. DOI: 10.1080/03075079.2021.1896801

Abstract. Internationalization is impacting universities and changing their core missions. In turn, many western universities have adopted a business model approach to deal with opportunities and challenges posed to their missions by internationalization. Resulting from increased scrutiny from the public and policy makers on the ability of universities to efficiently utilize public resources to achieve institutional missions, there is a growing interest to analyze this development and its effects upon the university business model. The purpose of this paper is to examine and evaluate how internationalization within the university's mission impacts the university's business model. Using a sample of German universities, this study develops a unique, three-stage, mathematical analysis to investigate this connection. By determining the internationalization and overall efficiencies of each institution relative to the other institutions in the dataset, it is found that no direct correlation between the relative internationalization efficiency and overall institutional efficiency exists, while also evidencing the use of efficiency analysis in allocating resources for internationalization and overall university mission achievement. These results show that while the relative efficiency of internationalization may contribute to a university's overall relative efficiency, other components of the university business model may also play key roles in determining overall relative efficiency, and the interplay of these components should be investigated in the future research.

Keywords. Higher Education; Internationalization; Efficiency; Business Model; University Mission; Performance Evaluation