Advancing Data Envelopment Analysis (DEA) with a Focus on the Evaluation of Hospital Efficiencies

Kumulative Dissertation

Der Wirtschaftswissenschaftlichen Fakultät

Der Universität Augsburg

Zur Erlangung des Grades eines

Doktors der Wirtschaftswissenschaften

(Dr. rer. pol.)

Vorgelegt von

Sebastian Bernhard Kohl

(Dipl. informationsorientierte VWL)

Erstgutachter: Prof. Dr. Jens O. Brunner

Zweitgutachter: Prof. Dr. Susanne Warning

Drittgutachter: Prof. Dr. Axel Tuma

Vorsitzender der mündlichen Prüfung: Prof. Dr. Marco Meier

Tag der mündlichen Prüfung: 02.12.2019

Table of Contents

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

Contribution 1:

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.

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/s10729-018-9436-8

Status: Published in *Health Care Management Science*, category A.

Contribution 2:

Kohl, S., Brunner, J. O. (2019). Benchmarking the Benchmarks – Comparing the accuracy of Data Envelopment Analysis models in constant returns to scale settings.

Status: Under revision ("revise and resubmit", May 15, 2019); submitted February 2, 2019. *European Journal of Operational Research*, category A.

Contribution 3:

Kohl, S. (2019). Using Data Envelopment to Estimate Hospital Efficiencies – A Teaching Case.

Status: Under revision; submitted June 28, 2019. *MSOR Connections*, no category available.

Acknowledgement

Special gratitude goes to my academic advisor Prof. Dr. Jens O. Brunner: I am very thankful for all the support and freedom, the inspiring and reassuring discussions, and the trust you showed in me during my PhD studies.

Moreover, I want to thank my colleagues at the chair for fruitful discussions and the provision of help, whenever I needed it. Thank you all for and a great time at the UNIKA-T.

I am very grateful for the support of my parents, who encourage me in every path I take! Thanks as well to my friends, especially those, who helped me to reach the finish line in windy conditions.

Thank you all very much.

1 Introduction and motivation

The importance of the health sector for the well-being of a population is undoubted. The economic dimension of the health sector, on the other hand, is often underestimated. This is especially true for the case of Germany, as it is one of the countries with the highest expenditures on healthcare in the world. With 11.3% of the GDP (Gross Domestic Product), Germany is ranked third in the worldwide comparison of health spending behind the United States of America and Switzerland (OECD 2017). Considering only the government/compulsory expenditures as a share of the GDP, Germany is even ranked first. The federal statistical office of Germany reported recently that for the first time in history, the daily health expenditures in Germany had passed the value of 1 billion Euros (Federal Statistical Office of Germany 2018). Issues in healthcare are high on the agenda of newspaper reports and the political discussion, as problems in the nursing sector currently show. Many hospitals and other healthcare institutions are struggling to cover the demand for nursing care. Nurses, on the other side, complain about bad working condition and low wages. Furthermore, the German government passed a law, restricting the number of patients a nurse is allowed to be responsible for, to improve the provision of patients with nursing care (Federal Ministry of Health 2018). All these circumstances will force affected institutions to spend more money on nursing care, causing a further increase of short to medium term expenditures. Due to an aging population as a result of the demographic change, the importance of the healthcare sector will increase additionally in the long run (Federal Statistical Office of Germany 2019b).

Within the healthcare sector, hospitals take a vital position. They are responsible for up to 40% of the healthcare expenditures in a country (OECD 2017). On the supply side, the pressure for hospitals to work cost efficient has increased significantly within the last 20 years. The introduction of the DRG (diagnosis related groups) system is responsible for a considerable part of the effect (Geissler et al. 2011). Hospitals now receive a fixed rate per case. Before the DRG introduction, all the costs of a patient's stay had been covered. As a result, the length of stay of inpatients has decreased significantly from 9.2 days in 2000 to 6.7 days in 2017 (Federal Ministry of Health 2019). More and more hospitals are closing or merging due to the increased cost pressure. While 2,242 hospitals existed in 2000, the number shrank to 1,942 in 2017 (Federal Statistical Office of Germany 2019a). This effect is evaluated diversly. While some fear for the supply of rural areas, others regard a thinning of the oversized German hospital sector as overdue (Busse & Berger 2018).

Looking at the demand for hospitals in Germany, a substantial rise in hospital admissions is visible. One of the main reasons is the changing age structure of the population (Krämer & Schreyögg 2019). Besides, the emergency departments of hospitals struggle with a growing share of outpatient cases that do not need emergency treatment (Scherer et al. 2017). These patients appreciate the immediate treatment possibilities without the need to apply for an appointment. This behavior worsens the situation of the hospitals as these cases cause congestion in the emergency departments while refunding is relatively low. Strategies have been developed to induce these people to use primary care resources instead (Köster et al. 2016).

Overall, the described situation causes pressure for hospitals to work efficiently. The latest edition of Health at a Glance: Europe (OECD/EU 2018) underlines the actuality of the topic. One of the two thematic chapters addresses "Strategies to reduce wasteful spending: Turning the lens to hospitals and pharmaceuticals". They argue that evidence suggests that up to one-fifth of health spendings is caused by inefficiencies and can be reduced without the performance of the health system. Therefore, the assessment of efficiency and identification of causes for inefficiency is an unavoidable first step. Furthermore, the identification of best practice examples is helpful for the adaption of commendable structures and processes. The scientific literature differentiates between two basic approaches for the estimation of efficiency: Parametric and non-parametric approaches (Jacobs et al. 2006)

While parametric approaches need the specification of the production process in a functional form, non-parametric methods can treat it as a black box. As a consequence, non-parametric models are superior in settings with multiple outputs and have prevailed in the estimation of hospital efficiencies (Hollingsworth 2003). The most popular method within the non-parametric approaches is the Data Envelopment Analysis (DEA). It is based on the theories of Farrell (1957) and was originated by the seminal paper of Abraham Charnes, William Cooper, and Edwardo Rhodes (Charnes et al. 1978). The basic CCR model has been named after their initials. The idea of the model is quite simple. Efficiency is expressed by the sum of weighted outputs divided by the sum of weighted inputs. The weights are decision variables in the resulting optimization model and necessary for several reasons. They assure flexibility and fairness of the model, as its most favorable set of weights evaluates every decision making unit (DMU). In doing so, no DMU can complain that its performance is the result of unfair weighting. Furthermore, the system of complete weight flexibility allows the mixture of highly diverse inputs and outputs, which do not have to be measured on the same scale or even in the same unit. The resulting efficiency term is maximized for the DMU under observation and yields the DMU's efficiency score. The constraint for the optimization problem restricts any DMU in the data sample to receive an efficiency score larger than 1 with the weights of the DMU under observation. Furthermore, weights are restricted to non-negative values. The optimization problem is given in (1).

Parameters & Sets:

n	Number of DMUs
m	Number of Inputs
s	Number of Outputs
$j = 1, \dots, n$	Set of DMUs with index j

$i = 1, \dots, m$	Set of inputs with index i
$r = 1, \dots, s$	Set of outputs with index r
$o \in 1, \dots, n$	DMU under observation
x_{ij}	Input i of DMU j
y_{rj}	Output r of DMU j

Decision variables:

v_i	Weight for input i
u_r	Weight for output r

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

$$\frac{\sum_{r=1}^{s} u_r y_{rj}}{\sum_{i=1}^{m} v_i x_{ij}} \leq 1 \qquad \forall j \tag{1b}$$

$$v_i, u_r \geq 0 \qquad \forall i, r \tag{1c}$$

A problem of (1) is its solvability, as decision variables appear as well in the nominator and denominator of (1a) and (1b). Therefore, (1) is a nonlinear problem and cannot be solved by the simplex algorithm. To alleviate the solution process, Charnes et al. (1978) used the Charnes-Cooper transformation (Charnes & Cooper 1962), which linearises the problem. They get rid of the denominator in (1a) by fixing it to 1 and adding it as a new constraint. Besides, (1b) is rewritten slightly. The results yield the typical CCR formulation (2).

$$\max \sum_{r=1}^{s} u_r y_{ro} \tag{2a}$$

$$\sum_{r=1}^{s} u_r y_{rj} \leq \sum_{i=1}^{m} v_i x_{ij} \qquad \forall j \tag{2b}$$

$$\sum_{i=1}^{m} v_i x_{io} = 1 \tag{2c}$$

$$v_i, u_r \geq 0 \qquad \forall i, r \tag{2d}$$

To receive an efficiency score for every DMU of a data sample, the mathematical program (2) needs to be solved n times, with every DMU being once the DMU under observation o. All efficient DMUs form a convex efficiency frontier, which "envelops" all other data points. This property was name-giving for the Data Envelopment Analysis.

In addition to the efficiency score, an optimal production plan for every DMU is created by DEA. For inefficient DMUs, this plan shows the changes they need to implement in order to become efficient. The DEA model obtains this plan from a linear combination of efficient reference units. This thinking is palpable in the dual problem of (2) (Charnes et al. 1978). The orientation of the optimal production plan can be modified with the mathematical program. (2) represents the CCR model with input-orientation. Input-oriented DEA models provide information on how to remove inefficiency by reducing inputs while keeping outputs constant until the DMU under observation is located on the efficient frontier. Output oriented models work the other way round. They expand the outputs until the DMU is located on the frontier until while holding all inputs at the initial level.

Beginning with the invention of DEA in 1978, a steady increase of publications in the field is visible. This trend has gained additional pace in the last ten years and shows signs of exponential growth (Emrouznejad & Yang 2018). Emrouznejad & Yang (2018) report the enormous number of 10,300 DEA related published journal articles, in total. Currently, around 1,000 new research articles are published every year. Since its foundation, the DEA methodology has been developed in a multitude of directions. While the basic CCR model is responsible for constant returns to scale (CRS) environments, a first model development (Banker et al. 1984) enabled the use of DEA in variable returns to scale (VRS) settings. This BCC model, which received its name again according to the initials of its founders (Rajiv Banker, Abraham Charnes, and William Cooper), is the second of the two basic models in DEA. The concept of constant and variable returns to scale modeling has been adapted to the vast majority of all later model developments. Further model developments try to overcome different shortfalls of the basic models or try to provide additional advancements. AR (assurance region) models, for example, restrict the weights to specific regions (Thompson et al. 1986). SBM (slacks-based measurement) models use the slacks of the dual LP to calculate the efficiency score (Tone 2001). Network DEA models (Färe & Grosskopf 2000) try to shed light on the transformation process from inputs to outputs, while fuzzy DEA models work with imprecise data (Kao & Liu 2000). These models are only some examples to illustrate the variety of DEA modeling. More information on existing DEA models can be found in Cooper et al. (2007).

Among the existing applications, Liu et al. (2013) identified the five areas, which are addressed most frequently. Their research shows the significance of DEA for healthcare, as it is the area with the second most publications. More DEA implementations are only available for the banking sector. Other popular fields of application are agriculture & farm, transportation, and education. According to Liu et al. (2013), within the healthcare field, the vast majority of publications is concerned with hospitals. Other healthcare applications deal with nursing homes, primary care, and care programs. Healthcare is not only one of the

sectors with the most applications, but also with the longest application history. The early adaption of DEA for healthcare indicates a natural fit of the method for the sector. The very first healthcare DEA study was conducted by Nunamaker (1983), who evaluated the performance of nursing services. The first hospital application (Sherman 1984) followed only slightly later. Nowadays, authors as Yasar A. Ozcan and Vivian Valdmanis are the leading contributors in the field. Added up, they published more than 70 research articles on the topic of healthcare DEA.

Some research questions drive this dissertation. In the beginning, an inventory should answer the questions:

1) What are the current developments in the field of healthcare DEA?
2) Which DEA models are famous in healthcare applications?

The first contribution addresses these questions. As the answers were as well surprising, as unsatisfactory, the main research question of this dissertation emerged:

3) What can be done to advance DEA in healthcare?

On purpose, the formulation of this central research question is very general. Therefore, not a single answer to the question exists. Instead, a pool of answers and suggestions is provided, which is influenced by all three contributions.

2 Summary of the contributions

This section discusses each contribution of this dissertation. The individual contributions are attached in the appendix.

2.1 The use of Data Envelopment Analysis (DEA) in healthcare with a focus on hospitals

As discussed in the last section, the dimensions of DEA research have reached an immense volume. To keep an overview of all publications is a physical impossibility. This emphasizes the need to divide the body of literature into subsections. The different fields of application seem a natural boundary for these subsections. The overwhelming body of literature, even within such a subsection, causes the need for additional guidance. In the field of healthcare applications, the literature reviews of Hollingsworth et al. (1999), Hollingsworth (2003, 2008), and O'Neill et al. (2008) help to gain an overview over the publications until 2004. Contribution 1 reviews the connecting period from 2005 to 2016 and closes a gap of over ten years of unreviewed publications. Furthermore, due to the increase in publications, the need for guidance in addition to the reprocessing of the existing publications is tangible. Contribution 1

addresses this need and acts as a roadmap for lessons learned regarding the conduction of DEA in the healthcare sector, especially for hospital applications.

Contribution 1 includes 262 publications from peer-reviewed journals in the review. As for the whole DEA literature, an increase in healthcare publications over the period under review is visible. Interesting trends are apparent when looking at the geographical focus of the studies. While studies regarding North America tend to stay at a constant level, European and Asian studies are on the rise. Other than at the beginning of the century, the annual publication rate of both regions now exceeds North America's. Furthermore, studies on the African continent arrived on the scene during the last decade. Nevertheless, their scope stays far behind the other regions. Health Care Management Science, the Journal of Medical Systems, and Health Policy have been identified as the journals with the most publications in the subsection. Furthermore, the research questions scientist are addressing have been reviewed systematically for the first time. Four main research clusters become visible:

Cluster	Number of publications in the period under review
Specific Management Questions	100
Pure DEA efficiency analysis	99
New methodology	48
Effects of reform	36

Figure 1: Research question clusters identified by contribution 1

Most studies are concerned with specific management questions. Within this cluster, a variety of different research objectives can be found. Some are trying to examine if quality affects efficiency. Others investigate if certain ownership types foster efficient hospital management or if specialization has a positive effect on efficiency. A large fraction of the publications under review is merely concerned with the conduction of a DEA study. They are summarized in the cluster pure DEA efficiency analysis. They often apply DEA for the first time in a particular country or use a model, which has not been applied to healthcare settings before. All 99 studies of the cluster share the property that the efficiency estimation itself is the reason for publication and not a tool to answer other questions. A surprisingly high number of publications (48) is concerned with the development of new methodology for healthcare settings or use healthcare applications to demonstrate their developments. This high number of methodological advancements in the subsection underpins the relevance and presence of the healthcare field in DEA. Although it is the smallest of the four clusters, a considerable number of publications uses DEA to identify the effects of reforms in the healthcare sector. The count of 36 publications should be valued even higher, as the cluster is far more specific than the others are. Therefore, the identification of the consequences of reforms on the efficiency in healthcare can be regarded as the most pressing single research question in

healthcare DEA. Especially publications concerning new methodology and specific management questions have seen considerable growth.

Regarding the methodology of the current publications, contribution 1 identifies and reviews three areas: data selection, model specification, and subsequent techniques. In the first area, the most relevant inputs and outputs are discussed. Little surprising, beds, medical staff, and nurses are the most commonly used inputs, while outpatients, inpatients, and other measures for the number of cases are the outputs on top of the list. These measures are the basis for a meaningful hospital analysis. However, the review shows plenty of other input and output categories, which can be used as inspiration for future studies. Especially the inclusion of quality data has been identified as a relevant topic for future studies. The trend shows an increasing rate of publications using quality measure in their studies. Furthermore, more and more studies that did not include quality measures designate this fact as a drawback and possible improvement for future studies. The increasing availability of the corresponding data will clear the way in this direction. In general, 3.8 inputs and 3.2 outputs have been used on average for a DEA study over the period of review. While a slight increase in the number of inputs is visible, the average number of outputs is decreasing to the same extent.

Apart from reviewing the model settings, contribution 1 raises awareness for some common mistakes concerning data usage. Although widely known as a pitfall, a significant amount of publications still uses an insufficient number of DMUs with regard to the number of inputs and outputs in their study. Different rules on the subject can be found in the literature. Dyson et al. (2001) advocate for the use of at least $2 \cdot (\#\text{inputs} + \#\text{outputs})$ DMUs in order to ensure sufficient discrimination between the units. However, 19 of the studies under review neglected this rule. Another problem in DEA applications concerning the data setup is the mixture of absolute and relative measures (Dyson et al. 2001). Distortion arises because absolute values depend on the size of a unit, while relative values share the same level for all units, independent of their size. In hospital applications, the issue mainly occurs, when bed occupancy rates or mortality rates are included in a study.

With regard to the model selection, contribution 1 found quite surprising results: Despite huge efforts in model development, the basic CCR and BCC models are by far the most utilized models. This seems even more surprising, as issues of these models like the missing integration of slack values into the efficiency score, or the assignment of zero weights are known for a long time.

The use of subsequent techniques is widely spread in DEA. Over two-thirds of the reviewed publications apply one of the techniques. With the term subsequent techniques, we summarize all methods that process DEA efficiency scores. The most common subsequent techniques are the Malmquist index, regression, and bootstrapping. While the use of the Malmquist index is slightly declining, the usage of regression and the bootstrapping has grown significantly. Regression as a second stage analysis is already in use for some time. Bootstrapping of DEA scores, on the other hand, is a more recent development. The publications of Simar & Wilson (2000a, 2000b, 2007) are mainly responsible for the trend to bootstrap DEA scores. The procedure provides two benefits: It enables the calculation of confidence intervals for DEA scores and

allows for the calculation of bias-corrected estimates. A bias of DEA scores is known to arise because the data at hand cannot entirely describe the true (unknown) efficient frontier. Efficient units might be missing in the data set, or no DMU has realized theoretically possible efficient production plans in reality. An overestimation of the frontier by all DEA models is the result (Bogetoft & Otto 2011). Because of these benefits, the bootstrapping procedure has become the most significant methodological trend of the last decade.

In order to enhance the quality of DEA studies in healthcare and to keep track of guidelines and other helpful publications, contribution 1 serves as a roadmap for lessons learned. Books and publications aside from the reviewed period or general publications outside the healthcare sector are as well considered. Through the combination of healthcare specific and unspecific publications, researchers find beneficial guidance for the conduction of state of the art DEA studies in healthcare. Finally, contribution 1 discusses relevant topics for future publications. The transformation of DEA from an almost exclusively scientific tool to an accepted instrument, which is deployed in practice, should be the ultimate target for the research community. The development of advanced models might be helpful in this regard. However, as long as the reliability of the results can be questioned, even the most advanced model will not be able to earn a sufficient amount of acceptance. Two ways are imaginable to foster confidence in DEA. The development of a procedure to prove the correctness and reliability of the results would be the most promising way. Another possibility lies in the further investigation of the results. A direct implementation of DEA results, as the calculation of a “projected DMU” or “improved activities” might pretend, is usually not applicable. No unit, independent of the economic sector, can cut down all relevant resources by a significant amount and keep the present output level. However, DEA is able to generate valuable insights by the identification of best practice examples and pointing at resources that are primarily responsible for inefficiency. Studies reporting additional profound process analyses about these resources might help to extend the understanding of DEA results. In addition, publications reporting the opinion of various experts on the results of a DEA study and their utilization possibilities would be of high value for the scientific community and might establish trust in DEA results for practitioners.

2.2 Benchmarking the Benchmarks – Comparing the accuracy of Data Envelopment Analysis models in constant returns to scale settings.

Despite the considerable effort spent on model development, the basic models are still those, which are applied most commonly in DEA. The CCR model (Charnes et al. 1978) is the basic model for constant returns to scale settings (CRS). Its counterpart for variable returns to scale (VRS) is the BCC model (Banker et al. 1984). The main reason for the high popularity of these models in applications is the lack of knowledge for more suitable options. The purpose of contribution 2 is to fill this gap and provide a benchmark for constant returns to scale settings that makes the accuracy of DEA models comparable. Based on this benchmark, contribution 2 evaluates the performance of different models and assess, if the

predominant position of the CCR model is justified. In doing so, the CCR model has to compete against the SBM (Tone 2001) and AR (Thompson et al. 1986) models. Furthermore, the BCC model and a random number generator are evaluated. The performance of these two models is expected to be inferior, as they are no suitable options for the evaluation of CRS data. Nevertheless, useful insights for the validation of the benchmarking procedure can be drawn from them.

The underlying idea of the procedure is to generate artificial data based on a sophisticated data generation process (DGP). For the generation of complete data samples, a Monte Carlo procedure is applied. The DGP utilizes a production function to generate meaningful input and output data for every DMU. Furthermore, the true efficiency score for every DMU is obtainable. The true efficiency score, which is unknown in real life applications, is a vital aspect of the procedure. It allows running a DEA with the input and output data in order to perform a comparison between the generated true efficiency scores and the efficiency scores estimated by the DEA model. A sophisticated DGP is essential for obtaining robust and trustworthy results and the derivation of general guidance. The existing literature in the field exposes several shortcomings in this regard. A sophisticated DGP consists of three key features.

First, a suitable production function is necessary. Most researchers in the field applied a Cobb-Douglas production function. However, due to its limited flexibility, it has been replaced by the Translog production function as state of the art over recent years. Contribution 2 develops the applicability of the Translog production function by providing a way for its utilization in pure CRS settings.

Second, the inclusion of a variety of characteristics, e.g., the number of inputs or the substitutability of inputs, is essential. These characteristics allow for the generation of different production scenarios. Almost all previous publications considered only two to four different characteristics. Contribution 2 uses eight different characteristics to ensure the results are not only valid for very particular production environments. The existing literature is extended in particular with regard to substitution effects. Contribution 2 develops two characteristics to adjust different settings on the substitution between inputs.

Finally, employing a sufficient number of levels for each characteristic is essential. In doing so, the levels should cover meaningful properties. For example, implementing the *number of inputs* characteristic with the levels *one* and *two* does not adequately reflect the majority of production processes, as these numbers are too small. Hence, the levels *three*, *five*, and *seven* are utilized in contribution 2, to represent a small, midsized, and large production process. Similarly, contribution 2 implements appropriate levels for all characteristics. Setting all characteristics to a certain level results in a specific production scenario. In combining all levels of all characteristics, contribution 2 employs 1,296 different scenarios. This multitude is a significant enhancement of the present literature, as no publication used more than 200 scenarios, so far.

In order to evaluate the scenario results, five different performance indicators are developed. They all compare the true efficiency scores with the estimates of the DEA model under evaluation. Every performance indicator covers a different area. Contribution 2 assess the ability of a DEA model to

1. minimize the deviation between the true efficiency score and the estimate
2. reproduce the correct ordering of the DMUs concerning their efficiency
3. identify efficient units
4. identify inefficient units
5. hit a corridor close to the true efficiency score with the estimate

A suitable efficiency estimator should be able to convince in all five dimensions. In order to receive a single benchmark for the accuracy of a DEA model, the average over the five dimensions is calculated. The resulting benchmark is called B.-Score and allows the comparison of the performance of different DEA models.

The robustness of results is a crucial issue, as the result of a scenario is stochastic. Therefore, the execution of replications is necessary. A scenario's final result is the average over all performed replications. Almost all previous studies use a fixed number of replications. Contribution 2 develops a flexible stopping criterion to take account of the increased computational effort due to the extension of scenarios while ensuring the robustness of the results at the same time.

The comparison of the five models with the benchmarking procedure shows an unambiguous dominance of the SBM and AR model over the basic CCR model. As expected, the BCC and random number generator perform worse than the remaining models. The characteristics having the largest effect on the results are the *number of DMUs* and the *number of inputs* of a scenario. Using more inputs leads to a decrease in the accuracy of the models. On a similar page, fewer DMUs lead to a deterioration in the accuracy of the models. Both observations are valid for all models analyzed in contribution 2. However, the SBM and AR models present themselves as less vulnerable in adverse scenarios. With regard to all 1,296 scenarios, both models are more robust than the others. Especially the reduced weight flexibility of the AR model seems to decrease the volatility of results.

When analyzing the influence of the number of inputs in more detail, a significant impact of the input correlation characteristic becomes visible. In general, the higher the correlation, the better the accuracy of the DEA models. While only a slight effect of the input correlation in an isolated view is visible, the situation changes significantly, when conducting a combined analysis with the number of inputs. For higher numbers of inputs, the accuracy of all models drops considerably, if no correlation between the inputs exists. Reassuring is the finding, that a slight correlation of 0.35 between the inputs is sufficient to mitigate the effect substantially.

Looking at the performance indicators, the overall results find support in every single performance indicator. Even on this level, the CCR performs always worse than the AR and SBM models. The inferiority of the CCR model in our overall result is therefore not depending on a single shortcoming or indicator definition, but visible over all relevant areas.

As contribution 2 identifies the number of DMUs and the number of inputs as characteristics with the largest impact on the accuracy of a DEA, an investigation of the prominent rule of thumb regarding the

usage of inputs and outputs is apparent. Several authors (i.a. Golany & Roll 1989, Dyson et al. 2001, Cooper et al. 2007) emphasized the importance to keep the ratio of DMUs to inputs and outputs at a reasonable level and formulated different rules of thumb. All these authors reason that disregarding the issue leads to insufficient discrimination and correctness of DEA results. Cook et al. (2014) argue that a statistical background for these rules is missing and their implementation often occurs out of convenience. Contribution 2 performs a separate study to elaborate on the subject. This study evaluates in succession different input levels. For every input level, the number of DMUs is examined, which is necessary to reach a predefined B.-Score with the CCR model. This predefined B.-Score represents the minimum accuracy a DEA study should achieve. The study is based on the property of the B.-Score to rise continually with higher DMU numbers. Therefore, an examination of an exact number of DMUs, which is necessary to reach the predefined value, is possible. With the results, contribution 2 can provide the background and reasoning for the utilization of a rule of thumb regarding the use of sufficient DMUs with regard to the number of inputs and outputs. Furthermore, contribution 2 shows that the existing rules of thumb underestimate the number of DMUs, which are necessary to conduct DEA studies of sufficient accuracy. In addition, the results do not support the linear dependency between the number of DMUs and inputs, which the existing rules of thumb assume. Therefore, contribution 2 suggests a new rule of thumb, representing the findings of the study.

2.3 Using Data Envelopment to Estimate Hospital Efficiencies – A Teaching Case

The conduction of meaningful Data Envelopment Analysis studies contains more pitfalls than many researchers are expecting. Contribution 3 provides a teaching case with a hands-on learning experience on the topic of hospital DEA. The teaching case addresses relevant issues regarding data, models, and result improvement. The data set of the case study is based on mid-sized German hospitals and provides a data set containing 70 DMUs. For every DMU, the number of beds, physicians, and nurses as inputs and the number of inpatients and outpatients as outputs are provided. These measures portray the service process of a hospital and are typical for hospital DEA studies. In addition, the Case Mix Index (CMI) and seven quality measures are included for every hospital in the data set. The CMI represents the average case severity of a hospital. It is a standard procedure to adjust the patient cases by the CMI to include the case severity in the analysis. Overall, contribution 3 leads through the process of conducting a DEA study with 15 questions, divided into the sections DEA modeling, data description, and results. While the teaching case mainly addresses junior DEA users, some parts are also of interest for more advanced DEA practitioners. Especially the implementation of the bootstrapping algorithm (Simar & Wilson 2000a) has to be highlighted in this context. The bootstrapping approach is a subsequent method that uses resampling techniques for result verification. It allows the calculation of bias-corrected efficiency scores. Although the inclusion of bootstrapping in DEA studies is more relevant than ever before, researchers often struggle

with the implementation, as replicable examples are rare and the existing publications on the subject are not always straightforward.

With regard to the utilized DEA models, the conductor of the case study is concerned with the CCR and SBM model. In doing so, the impact of weights and slacks for the analysis is brought into focus. In this way, the vast number of zero weights, the CCR model assigns, becomes apparent. Furthermore, the super-efficiency model finds utilization in the detection of data outliers. The exclusion of these outliers is essential, as otherwise, the results of the study are likely to be distorted. Besides the outlier detection, the data section addresses as well the treatment of missing values. This issue is widespread for the conduction of DEA studies, as a comprehensive data set hardly ever exists.

Another central component of the case study is the inclusion of quality measures in DEA. In the data section, the conductor of the case study gets to know the different dimensions of hospital quality. The topic is as well part of the modeling section. A rarely applied two-stage procedure is used to include the quality indicators into DEA. In this two-stage procedure, the Helmsman DEA is used in the first stage to create a single quality indicator out of the quality measures at hand. The single quality indicator is then multiplied with the patient cases in order to weight the case number with the quality of the hospital. In the second stage, a regular DEA with the adjusted measures is conducted. The approach has the benefit to prevent the DEA model from excluding quality completely by assigning zero weights. This often occurs, when quality indicators are used as additional outputs in a DEA study. Furthermore, the procedure is not increasing the number of inputs and outputs in the main DEA study. This is in line with the advice of Dyson et al. (2001) to be parsimonious with the number of inputs and outputs.

3 Discussion of the contributions

In this section, the research questions posed at the beginning of this dissertation are discussed. The findings of all contributions are brought together to answer these questions.

Research question 1: What are the current developments in the field of healthcare DEA?

Finding the answer to this question is one of the main goals of contribution 1. A first finding is that the trend to apply DEA in healthcare settings, especially in hospital environments, is undaunted. Similar to the overall trend, a significant increase in DEA studies related to the healthcare field is visible. This fact represents a basis for the relevance of this thesis. The geographical origin of the studies, however, changed significantly. The annual number of publications from Europe and Asia now exceeds the one from North America. Furthermore, the African continent is no longer a blank spot for healthcare DEA studies.

The examination of the reasons for researchers to conduct healthcare DEA studies exposed two main trends: The analysis of specific management questions and the development of new methodology. The increased use of DEA for the examination of specific management questions is a good sign. It documents the variety of specific fields of application for healthcare DEA and the need for a multicriterial benchmarking method. One of the main subjects in this regard is the effect of quality. These studies explicitly investigate whether the quality of provided services affects as well the unit's efficiency. Furthermore, more and more researchers care about the inclusion of quality indicators in healthcare DEA studies. The ongoing development of the methodology is driven by the apparent shortcomings of the basic models. Although a vast body of publications exists, the doubts in the results of DEA have not vanished. Another development in the field is the utilization of the bootstrapping method in applications. It serves as a tool for bias correction and the calculation of significance intervals for the results. The popularity of the bootstrap is a further expression for the desire to improve the DEA methodology and provide consistent and robust results. Contribution 1 reports a significant increase in the usage of the bootstrap between 2005 and 2016. Finally, a trend to use regression in a second stage of the study is apparent. In doing so, researchers try to understand the coherence of efficiency with environmental factors. Finding the factors that mainly affect a DMUs efficiency allows the provision of more detailed guidance for managers or politicians.

Research question 2: Which DEA models are famous in healthcare applications?

As shown in research question 1, the development of models is one of the primary research purposes in the field. This leads to the question, which models healthcare DEA applications utilize most frequently. The insights of contribution 1 on this issue are quite unexpected. The models with by far the most applications are the basic CCR and BCC models. At the first moment, this finding might seem natural, as the CCR and BCC models are the embodiment of DEA. On second thought, the finding is astonishing, considering the efforts spent on model development for over 40 years. However, as no other model has

gained acceptance as a new standard, researchers still apply the basic models. Especially surprising is the gap in utilization, in comparison to all other models. The model utilized most after the basic ones, is the super-efficiency model. It has been applied in 14 out of 262 studies, while the CCR model records 112 and the BCC model 144 applications.

Almost 80% of studies utilize at least one of the two models. Quite rightly, some compare the results of new model developments with the established ones or use multiple models for other reasons. Nevertheless, almost 65% of the publications use only the basic models for the efficiency estimation of their studies. One in four studies is entirely limited to the use of a basic model. Not even subsequent techniques or second stage analyses as Malmquist indices, regression, or bootstrapping find application in these studies. These findings express a deficiency of DEA, as essential developments do not access the actual applications.

Research question 3: What can be done to advance DEA in healthcare?

Research question 3 is the core question of this dissertation and the answers are manifold. All three contributions are used to provide answers to this question.

First, it is essential to stop study misspecifications, where approved rules are existing. By addressing these issues in a literature review (contribution 1) and a teaching case (contribution 3), they are made visible and distributed for a broad audience. Examples are the mixture of absolute and relative data or the usage of an insufficient number of DMUs, considering the number of inputs and outputs. Both problems are known to lead to a distortion of results (Dyson et al. 2001). As studies with these issues get published, even reviewers seem unaware of the problems. Contribution 1 and contribution 3 especially raise the awareness of junior DEA researchers to these issues, as literature reviews and teaching cases are natural starting points into a research area. Contribution 3 also addresses the identification of data outliers and the treatment of missing data. The neglection of these topics leads as well to avoidable study misspecifications.

Second, many helpful guidelines for the conduction of DEA in general, as well as for healthcare DEA, are existing. These encompass lessons learned for several topics, as model usage or the identification of suitable inputs and outputs. The vast body of literature is complicating the knowledge about all these valuable publications. Especially for researchers, who are new to the field, this presents a tough challenge. To gather the information on existing guidelines, contribution 1 provides a roadmap to lessons learned.

Third, contribution 1 dedicates a whole section on essential topics for future publications. The enhancement of DEA study quality is the core issue of this section and should raise awareness for the subject. The section addresses two particular paths to foster confidence in DEA studies. One is the creation of a method to assess the accuracy of DEA models. The other is to prove the validity of results by documenting successful result implementations in real cases. By addressing these issues, further research might dedicate more effort on the enhancement of healthcare DEA and pursue these two directions.

Fourth, more effort is necessary to represent the service process of hospitals more precisely. The meaningful inclusion of quality indicators in all hospital DEA studies is essential in this matter. Contribution 1 shows that some authors already try to proceed in this matter. However, the share of hospital

studies including quality still needs to grow significantly. More importantly, methods that ensure the actual consideration of quality measures need to prevail. Most common is the inclusion of quality indicators as additional output. However, most DEA models can erase the indicator for units with a poor quality performance from the analysis. As a result, using quality parameters as additional outputs can never lead to a decline in efficiency scores. The results of such studies are highly misleading, as DMUs with poor quality might arise with higher scores than without the inclusion of the quality output. Spreading a method, which overcomes this obstacle is one of the goals of contribution 3. The promoted two-stage procedure has two major benefits. First, it allows the calculation of a single quality indicator out of various measures via the helmsman method in the first stage. This facilitates the inclusion of different quality dimensions, while at the same time, the number of outputs is not inflated. Second, the procedure prevents a neglection of quality in the final score through a multiplication of the single quality indicator with the case numbers in the second stage.

Fifth, the usage of the bootstrapping method raises the quality of DEA studies, as it allows for the calculation of confidence intervals and bias correction of results. Contribution 1 promotes the method by showing its increased relevance in healthcare DEA studies. Furthermore, contribution 3 demands the implementation of the bootstrap in the teaching case. By providing an example with a comprehensive explanation of the solution and the corresponding implementation, the hurdle for the use of bootstrapped DEA is lowered. So far, comprehensible examples on the implementation of the bootstrap are rare.

Sixth, contribution 3 fosters the understanding of the DEA mechanics in comparing the weight assignment of the CCR and SBM models in detail. With more researchers being aware of the differences, more reasonable conclusions will be drawn from DEA studies. Besides, highlighting the issues of the basic models might encourage more researchers to use advanced models.

Seventh, as concluded in contribution 1, a method to assess the accuracy of DEA models would be extremely beneficial in advancing the DEA methodology. Contribution 2 provides precisely such a system. The invention of the B.-Value enables the judgment of a model's accuracy based on a single score. In addition, the development of the B.-Rank in contribution 2 facilitates the comparison of models with similar B.-Values. Therefore, the methodology allows the comparison of the accuracy of existing DEA models. Furthermore, researchers inventing new model developments can prove their reasonability with the help of the method. For a start, contribution 2 shows that the CCR results are less accurate in constant returns to scale settings, than these of the SBM and AR model. Due to a slight dominance of the SBM in the B.-Rank and more natural feasibility, the use of the SBM as the new gold standard for DEA is advocated.

Eighth, contribution 2 provides valuable general insights for the applicability of DEA. Delivering information, under which circumstances DEA grants results of satisfying accuracy is a major contribution to the advancement of future studies. In this regard, especially the usage of a sufficient number of DMUs depending on the number of inputs and outputs is of interest. Contribution 2 shows that existing rules on the subject do not assure adequate study quality as the supposed number of DMUs is too low. Furthermore,

the linear dependency, existing rules assume, is not supported by the study results. This amplifies the issue in settings with larger numbers of inputs and outputs. Contribution 2 develops a new rule to overcome these shortcomings. Adherence to this rule will advance the quality of future DEA studies significantly.

4 Conclusion

This dissertation is concerned with the advancement of data envelopment analysis, especially for the implementation in hospital settings. It comprises three contributions: The first contribution is a literature review, which fills an unreviewed gap of over ten years and encompasses 262 publications. The second contribution develops a methodology to assess the accuracy of DEA models. The third contribution sets up an advanced teaching case to distribute knowledge on the DEA methodology among researchers and practitioners. After a motivation of the topic at the beginning of this theses, three research questions are derived that guide the dissertation. The findings of contribution 1 concerning research questions 1 and 2 lead the way towards the central topic of this thesis, which is stated in research question 3:

What can be done to advance DEA in healthcare?

The answers to this central question comprise findings of all contributions. Most importantly, the development of a method to compare the accuracy of DEA models is an essential step for the advancement of DEA. For the first time, different DEA models can be compared and evaluated based on a neutral criterion. The benchmarking method allows the judgment of existing models, as well as the trial of new model developments. This procedure allows the formation of new gold standards in DEA, which can replace the basic CCR model. An experimental study in the second contribution shows that the SBM outperforms the CCR model and should be the new standard for the evaluation of constant returns to scale studies. This finding is generally valid for all DEA applications. Its particular relevance for healthcare DEA is supported by the answer to research question 2, which highlights an inadequate representation of sophisticated models in healthcare applications.

One of the answers to research question 1 reveals another main field for advancing healthcare DEA studies: A trend to include quality measures into the analysis has started. The contributions 1 and 3 promote the topic, in order to raise the share of studies considering quality in their analysis. Besides a rising share of studies considering quality, meaningful implementation of quality indicators into the studies is of prime importance. Contribution 3 presents a two-stage approach using a Helmsman DEA in the first stage. Its benefits lie in a suitable inclusion of multiple indicators. Furthermore, it assures that the DMUs cannot evade the evaluation of quality. In doing so, the inclusion of quality parameters via the Helmsman DEA method is not automatically increasing the average efficiency in a data sample, as other conventional methods do. Finally, the method is parsimonious with inputs and outputs, which supports another core finding of this thesis: The adherence to meaningful settings for DEA is without any alternative. In this regard, the setting with the most substantial impact is the restrainment of inputs and outputs with regard to

the available DMU number. Contribution 2 underpins the influence of this setting on the accuracy of results and shows that existing guidelines cannot guarantee a sufficient estimation accuracy. Therefore, a new guideline is developed, which ensures a higher quality of results for future studies.

The contributions of this thesis open plenty of possibilities for future research. First of all, the presented benchmarking procedure for the accuracy of DEA models focusses on constant returns to scale settings. An extension of the method to variable returns to scale settings is a natural next step. Furthermore, the models under assessment in the experimental study of contribution 2 are a first selection. Further evaluations can reveal other existing models that outperform the SBM model. Examining a combination of the SBM and AR models, as suggested in Tone (2001) or testing for the best AR restrictions are as well suitable future research projects.
Besides, advancing the development and application of the bootstrap for other models than the CCR is desirable for future research. As the method evolves to become state of the art in DEA, a shift of the standard DEA model to more sophisticated models should not be prevented by missing bootstrapping procedures for these models.

In general, the contributions present encouraging results regarding the accuracy of DEA. Not every DEA model delivers satisfying results, and in some settings, DEA should not be applied at all. However, if sophisticated models are used in an appropriate environment, DEA is a valuable method for supporting decision makers. The contributions of this thesis are meant to advance DEA for healthcare settings and help the methodology to find acceptance and application in actual management support.

5 References

Banker, R. D., A. Charnes, W. W. Cooper (1984). Some Models for Estimating Technical and Scale Inefficiencies in Data Envelopment Analysis. *Management Science* **30**(9) 1078-1092.

Bogetoft, P., L. Otto. *Benchmarking with DEA, SFA, and R.* Springer New York. New York, NY:

Busse, R., E. Berger (2018). Weniger (Standorte, Betten und Fälle) ist mehr (Zugang, Qualität und Ergebnisse). Standpunkte der Gesundheitsökonomie. In *Krankenhauslandschaft in Deutschland: Zukunftsperspektiven-Entwicklungstendenzen-Handlungsstrategien*, Böcken, Bühler, Lasserre, Simic, Stock, Henriksen, Knieps, Zich, Jendges and Leber (eds.), Kohlhammer Verlag.

Charnes, A., W. W. Cooper (1962). Programming with linear fractional functionals. *Naval Research Logistics Quarterly* **9**(3-4) 181–186.

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

Cook, W. D., K. Tone, J. Zhu (2014). Data envelopment analysis: Prior to choosing a model. *Omega* **44** 1–4.

Cooper, W. W., L. M. Seiford, K. Tone. *Data envelopment analysis: a comprehensive text with models, applications, references and DEA-solver software*. Springer Science & Business Media. :

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

Emrouznejad, A., B. R. Parker, G. Tavares (2008). Evaluation of research in efficiency and productivity: A survey and analysis of the first 30 years of scholarly literature in DEA. *Socio-Economic Planning Sciences* **42**(3) 151–157.

Emrouznejad, A., G.-l. Yang (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) p 253-290.

Federal Ministry of Health. (2018, Oct 5). Verordnung zur Festlegung von Pflegepersonaluntergrenzenin pflegesensitiven Bereichen in Krankenhäusern. Pflegepersonaluntergrenzen-Verordnung - PpUGV. Retrieved from https://www.bundesgesundheitsministerium.de/personaluntergrenzen.html.

Federal Ministry of Health. (2019, Apr 12). Krankenhausfinanzierung. Retrieved from https://www.bundesgesundheitsministerium.de/krankenhausfinanzierung.html.

Federal Statistical Office of Germany. (2018, Feb 15). Gesundheitsausgaben pro Tag überschreiten Milliardengrenze. Retrieved from https://www.destatis.de/DE/Presse/Pressemitteilungen/2018/02/PD18_050_23611.html.

Federal Statistical Office of Germany. (2019a, Jun 21). Anzahl der Krankenhäuser in Deutschland in den Jahren 2000 bis 2017. Retrieved from https://de.statista.com/statistik/daten/studie/2617/umfrage/anzahl-der-krankenhaeuser-in-deutschland-seit-2000/.

Federal Statistical Office of Germany. (2010b, Jul 14). Prognostizierte Entwicklung der Altersstruktur in Deutschland von 2010 bis 2050. Retrieved from https://de.statista.com/statistik/daten/studie/163252/umfrage/prognose-der-altersstruktur-in-deutschland-bis-2050/.

Geissler, A., D. Scheller-Kreinsen, W. Quentin, R. Busse (2011). Germany: Understanding G-DRGs. In *Diagnosis-related groups in Europe. Moving towards transparency, efficiency and quality in hospitals,* Busse, Alexander Geissler, Wilm Quentin and Miriam Wiley (eds.), Maidenhead, England, Open University Press.

Golany, B., Y. Roll (1989). An application procedure for DEA. *Omega* **17**(3) 237–250.

Hollingsworth, B. (2003). Non-parametric and parametric applications measuring efficiency in health care. *Health Care Management Science* **6**(4) 203–218.

Hollingsworth, B. (2008). The measurement of efficiency and productivity of health care delivery. *Health Economics* **17**(10) 1107–1128.

Hollingsworth, B., P. J. Dawson, N. Maniadakis (1999). Efficiency measurement of health care: a review of non-parametric methods and applications. *Health Care Management Science* **2**(3) 161–172.

Jacobs, R., P. C. Smith, A. Street. *Measuring efficiency in health care: analytic techniques and health policy*. Cambridge University Press. :

Kao, C., S.-T. Liu (2000). Fuzzy efficiency measures in data envelopment analysis. *Fuzzy sets and systems* **113**(3) 427–437.

Köster, C., S. Wrede, T. Herrmann, S. Meyer, G. Willms, A. Seyderhelm, T. Seeliger, B. Broge, J. Szecsenyi (2016). Ambulante Notfallversorgung. Analyse und Handlungsempfehlungen. Göttingen: AQUA – Institut für angewandte Qualitätsförderung und Forschung im Gesundheitswesen GmbH.

Krämer, J., J. Schreyögg (2019). Demand-side determinants of rising hospital admissions in Germany: the role of ageing. *The European Journal of Health Economics*.

Lewis, H. F., T. R. Sexton (2004). Network DEA: efficiency analysis of organizations with complex internal structure. *Computers & Operations Research* **31**(9) 1365–1410.

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

Nunamaker, T. R. (1983). Measuring routine nursing service efficiency: a comparison of cost per patient day and data envelopment analysis models. *Health services research* **18**(2 Pt 1) 183.

O'Neill, L., M. Rauner, K. Heidenberger, M. Kraus (2008). A cross-national comparison and taxonomy of DEA-based hospital efficiency studies. *Socio-Economic Planning Sciences* **42**(3) 158–189.

OECD (2017). Health at a Glance 2017. OECD Indicators. *OECD Publishing, Paris.* http://dx.doi.org/10.1787/health_glance-2017-en.

OECD/EU (2018). Health at a Glance: Europe 2018. State of Health in the EU Cyle. *OECD Publishing, Paris.* https://doi.org/10.1787/health_glance_eur-2018-en.

Scherer, M., D. Lühmann, A. Kazek, H. Hansen, I. Schäfer (2017). Patients Attending Emergency Departments. *Deutsches Arzteblatt international* **114**(39) 645–652.

Sherman, H. D. (1984). Hospital Efficiency Measurement and Evaluation: Empirical Test of a New Technique. *Medical Care* **22**(10) 922–938.

Simar, L., P. W. Wilson (2000a). A general methodology for bootstrapping in non-parametric frontier models. *Journal of Applied Statistics* **27**(6) 779–802.

Simar, L., P. W. Wilson (2000b). Statistical Inference in Nonparametric Frontier Models: The State of the Art. *Journal of Productivity Analysis* **13**(1) 49–78.

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

Thompson, R. G., F. D. Singleton Jr, R. M. Thrall, B. A. Smith (1986). Comparative site evaluations for locating a high-energy physics lab in Texas. *interfaces* **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.

A. Appendix

A1. The use of Data Envelopment Analysis (DEA) in healthcare with a focus on hospitals

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.

The final authenticated version is available online at: https://doi.org/10.1007/s10729-018-9436-8

Status: Published in *Health Care Management Science*, category A.

A2. Benchmarking the Benchmarks – Comparing the accuracy of Data Envelopment Analysis models in constant returns to scale settings

Kohl, S., Brunner, J. O. (2019). Benchmarking the Benchmarks – Comparing the accuracy of Data Envelopment Analysis models in constant returns to scale settings.

Status: Under revision ("revise and resubmit", May 15, 2019); submitted February 2, 2019. *European Journal of Operational Research*, category A.

Benchmarking the Benchmarks – Comparing the accuracy of Data Envelopment Analysis models in constant returns to scale settings

Sebastian Kohl[a,b], Jens O. Brunner[a,b,*]

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

[b] University Center of Health Sciences at Klinikum Augsburg (UNIKA-T), Neusässer Straße 47, 86156 Augsburg, Germany

* Corresponding author: jens.brunner@unikat.uni-augsburg.de, +498125986446, ORCID 0000-0002-8793-955X

Abstract

Despite the massive use of Data Envelopment Analysis (DEA) models in scientific applications, no publication cared about identifying the DEA model, which is able to provide the most accurate efficiency estimates, so far. We develop an established method based on a Monte Carlo data generation process to create artificial data. As we use a Translog production function instead of the commonly utilized Cobb Douglas production function, we are able to construct meaningful scenarios for constant returns to scale. The generated data is then assessed by five different DEA models. Finally, the quality of the resulting efficiency estimates is evaluated by five performance indicators and summarized in benchmark scores. With this procedure, we can postulate general statements on parameters that influence the quality of DEA studies in a positive/negative way and determine which DEA model operates in the most accurate way for a range of scenarios. Here, we can show that the Assurance Region and Slacks-Based-Measurement models outperform the CCR (Charnes-Cooper-Rhodes) model in constant returns to scale scenarios. We therefore recommend a reduced utilization of the CCR model in DEA applications.

Keywords: Data Envelopment Analysis, Monte Carlo experiments, Artificial Data

1 Introduction

From the invention of DEA in 1978 (Abraham Charnes, William Cooper, & Rhodes, 1978) a continuous growth in applications and model developments can be explored (Emrouznejad & Yang, 2018). Nowadays, it is extremely hard to get an overview over the whole range of DEA models. As a comparison of the performance of different DEA models is missing in the current literature, it is difficult to choose a proper model for an application. A recent literature review in the field of healthcare (Kohl, Schoenfelder, Fügener, & Brunner, 2018) revealed a very predominant position of the basic models (CCR & BCC) in applications. On the one hand, this fact is surprising, as the basic models struggle with known problems as slacks and zero weights (WilliamW. Cooper, Ruiz, & Sirvent, 2011; Pedraja-Chaparro, Salinas-Jimenez, & Smith, 1997). On the other hand, it is the obvious implication of a lack of knowledge regarding options that are more suitable. At the bottom line, the uncertainty about the accuracy of the efficiency estimates of DEA models and the model that yields the best results (for certain circumstances) is the biggest issue of DEA. From our point of view, it is the crucial point that prevents DEA from leaving the scientific stage and finding actual application by politicians, economists, and managers.

A judgement of the quality of DEA estimates in real data applications however is not possible, as the true efficiency values are unknown. To overcome this obstacle, we generate artificial data, where the true efficiency of every decision-making unit (DMU) is known. Therefore, it is possible to compare the estimates from different DEA models with the corresponding true values and make statements on the quality of the models. In order to derive meaningful conclusions, two aspects for the generation of artificial data play a key role. First, a sophisticated Data Generation Process (DGP) is necessary to create reasonable data for the DMUs. Second, the consideration of a multitude of different scenarios is essential to derive generally valid results. By respecting both aspects, the generation of meaningful data is possible via a Monte Carlo Simulation process. Figure 1 depicts the whole process of our procedure. All parts of the process will be described in detail in the upcoming sections.

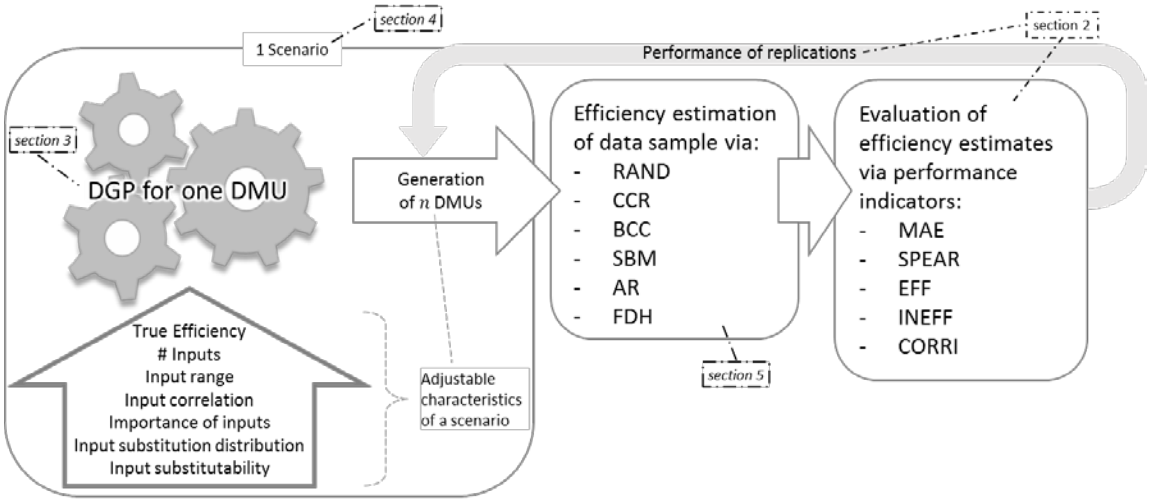


Figure 1 - The process of Benchmarking the Benchmarks for one scenario

Although no paper had the goal to identify the best DEA model so far, Monte Carlo simulated data has already been used to evaluate certain features of DEA. Table 1 gives an overview of these studies and shows main characteristics as production function used, number of replications, inputs, outputs etc. The identification of the best DEA model however, has not been the goal of a study, yet. Furthermore, both the DGP, as the scenario variety show significant room for improvement. As Table 1 shows, most papers employed only the very restricted Cobb-Douglas (CD) production function in their DGP (Cordero et al. 2015). Siciliani (2006) for example criticize the CD for its inflexibility regarding input substitution elasticity and scale effects. Over the last years, the Translog production function (TL) has emerged as a reasonable alternative. It is a generalization of the CD and allows for the creation of more realistic data. Therefore, we use the TL for the DGP in this paper. The number of inputs in nearly all studies is set too low, as the literature review of Kohl et al. (2018) exhibits a median of 4 inputs over 262 DEA applications in hospital settings. Furthermore, most studies only use one setting for the input number. In general, too little attention has been paid on the generation of a widespread range of scenarios. This is as well reflected by the observation, that most studies only change three or less characteristics in their DGP

Author	Production Function apart from CD	# Replications	# Inputs	# Outputs	# Scenarios	# Characteristics to generate scenarios	Adjustable Substitution Effects for CRS	Evaluated DEA models apart from CCR/BCC	Comparison with parametric models?
Banker et al. 1993	-*	5	2	1	128	4	No	-	Yes
Banker et al. 1996	-*	25	2	1	48	4	No	-	Yes
Pedraja Chaparro et al. 1997	-	≥6	1-6	1	24	2	No	AR	No
Smith 1997	-	≥125	2-6	1	20	2	No	-	No
Yu 1998	CRESH	25	3	1	21	2	Yes	BM	Yes
Zhang & Bartels 1998	-	5-100	2	1	80	3	No	-	No
Pedraja-Chaparro et al. 1999	-	≥100	2-20	1	190	3	No	-	No
Resti 2000	-	N/A	2	3	6	2	No	-	Yes
Holland & Lee 2002	-	1,000	2	1	24	3	No	FGK	No
Ruggiero 2007	-	100	2	1	20	2	No	RM	Yes
Van Biesebroek 2007	-**	50	2	1	9	3	No	-	Yes
Perelman & Santín 2009	TL	100	2	2	12	2	No	-	No
Krüger 2012	TL, CES	1,000	2	1	144	7	Yes (for CES)	FDH	Yes
Cordero et al. 2015	TL	1,000	3	1	28	3	No	-	No

* A piecewise CD is used ** CD is used as output constraint in a NPV maximization model over time

CRESH = Constant Ratio of Elasticity of Substitution, Homothetic; TL = Translog; CES = Constant Elasticity of Substitution

AR = Assurance Region; BM = Banker and Morey model; FGK = Färe, Grosskopf, Kokkelenberg model; RM = Russell measure; FDH = Free disposal hull

Table 1 - Literature of DEA evaluations via Monte Carlo simulated data

Most studies in the field tried to explore properties of the basic DEA models (Smith 1997, Zhang & Bartels 1998, Pedraja-Chaparro et al. 1999, Holland & Lee 2002). Very popular are as well comparisons between the basic DEA and parametric models. The consideration of DEA models apart from the basic CCR and BCC models is rather sparse. Only about one third of the studies is considering an alternative DEA model at all and none has considered more than one so far. The robustness of results is another issue with regard to previous papers. As the DEA estimators depend on the random data, the conduction of replications for each scenario is indisputable. Krüger (2012) criticizes in this context the low number of replications of many studies. The number of replications varies significantly between studies in a range from 5 to 1,000 (see Table 1).

The purpose of this paper is to fill the gap on the uncertainty of the accuracy of DEA models and provide a procedure to make the accuracy of DEA models comparable. We bring the reliability of DEA results into question and show that the environment of a DEA study has a high influence on the accuracy of the estimation results. Light is shed on the "blackbox" of DEA that should raise the awareness that its results

are estimates and should be treated accordingly. In doing so, we focus on constant returns to scale (CRS) settings and try to answer the question, if the predominant position of the CCR model is justified.

The contribution of this paper is manifold. First, we address the question, whether the CCR model should be further on used as the standard model for the evaluation of constant returns to scale settings. For this reason, we compare the CCR model with an Assurance Region (AR) and a Slack-Based Measurement (SBM) model. A comparison with uniformly distributed random numbers (RAND) and the basic model for variable returns to scale (BCC) provide further insight into the reliability and accuracy of the procedure. Second, we create a benchmark score that inherits multiple performance indicators. These performance indicators cover all relevant properties of an efficiency estimator as e.g. the identification of efficient units and the correct ordering of the units' efficiency values in a sample. Having one benchmarking score, we can show how the DMUs / (Inputs + Outputs) ratio has to be set, to achieve a satisfying quality for DEA applications. Hence, we derive a new rule for the minimum number of DMUs in DEA applications. Third, in order to address the problem of general result validity, we advance the scenario variation significantly. A scenario is always a concrete combination of values for all characteristics (e.g. number of DMUs, inputs generation characteristics, efficiency score). We generate 1,296 distinct scenarios, whereas no study had used more than 200 scenarios so far (see Table 1). Even more important than just the pure number of scenarios is the coverage of different characteristics, which influence the DGP and the creation of scenarios. With eight different characteristics, we provide another significant advancement compared to the existing papers. Fourth, we show how to use the Translog production function in CRS settings. This allows an adjustment of the input substitution, which leads to a more realistic DGP. So far, only Krüger (2012) and Yu (1998) looked at different input substitutions by the use of a Constant Elasticity of Substitution (CES) and Constant Ratio of Elasticity of Substitution, Homothetic (CRESH) production functions (see Table 1). The Translog production function however is an even more flexible production function and allows for additional settings on input substitution. Fifth, we divide the input substitution into the two distinct effects of substitutability and substitution distribution. This contributes as well to the generation of more realistic data, as to the variety of scenarios. Sixth, we turn to the problem of result robustness. We recognize that the necessary number of replications for each scenario highly depends on the number of DMUs. While 50 replications might be sufficient for scenarios with 450 DMUs, 200 replications might not be sufficient in a scenario with 50 DMUs. Therefore, we develop a flexible stopping criterion for the generation of replications.

The remainder of this paper is organized as follows: Section 2 depicts the applied performance indicators and the benchmarking procedure itself. In Section 3 we present our data generation process for one DMU in detail. The design of our study is described in Section 4 and Section 5 is concerned with the evaluation of our study. The results of our experimental study follow in Section 6. Finally, Section 7 provides a summary and conclusion.

2 Performance indicators & benchmarking procedure

As the purpose of this paper is the comparison and evaluation of different DEA models, the performance indicators for this evaluation are playing a key part. Pedraja-Chaparro et al. (1999) focused on identifying such performance indicators for Monte Carlo DEA analyses. Their results form the basis for our performance indicators. They argue, "any judgement on the quality of a DEA model must be made in the light of the purposes for which the results are used" and name four main purposes of a DEA:

1) Identifying inefficient DMUs,
2) Ranking the performance of DMUs,
3) Estimating efficiencies and setting targets for improvement, and
4) Examining the overall efficiency of an industry.

For our model evaluation, we utilize five different performance indicators. They cover all areas identified by Pedraja-Chaparro et al. (1999). Table 2 features all indicators, with θ_j denoting the true efficiency of DMU $j = 1, \dots, n$, whereas $\hat{\theta}_j$ is the score estimated by the DEA model. The true efficiency value is available for all DMUs through the DGP described in Section 3. All indicators are adapted to have one as the best possible value with lower values always being worse.

The mean absolute error (MAE) and the Spearman Rank correlation index (SPEAR) are the most obvious indicators. Their usage has already been established in former studies (Yu 1998, Krüger 2012, Cordero et al. 2015). They represent the expectation that estimates differ only to a small extent from the true values and the ordering of the efficiency estimates corresponds to the true ordering. They represent the purposes 4) *Examining the overall efficiency of an industry* and 2) *Ranking the performance of DMUs* listed above.

Indicator	Definition
MAE	$1 - \frac{1}{n}\sum_{j=1}^{n} \lvert\theta_j - \hat{\theta}_j\rvert$
SPEAR	$\frac{\sum_j \left(\mathrm{Rg}(\theta_j) - \overline{\mathrm{Rg}(\theta)}\right)\left(\mathrm{Rg}(\hat{\theta}_j) - \overline{\mathrm{Rg}(\hat{\theta})}\right)}{\sqrt{\sum_j \left(\mathrm{Rg}(\theta_j) - \overline{\mathrm{Rg}(\theta)}\right)^2}\sqrt{\sum_j \left(\mathrm{Rg}(\hat{\theta}_j) - \overline{\mathrm{Rg}(\hat{\theta})}\right)^2}}$
EFF	$\frac{\left\lvert\{j \in \{1,\dots,n\}: \theta_j \geq Q(\varepsilon) \cap \hat{\theta}_j \geq Q(\varepsilon)\}\right\rvert}{\left\lvert\{j \in \{1,\dots,n\}: \theta_j \geq Q(\varepsilon)\}\right\rvert} \cdot \left(1 - \frac{\max\{\left\lvert\{j \in \{1,\dots,n\}: \hat{\theta}_j \geq Q(\varepsilon)\}\right\rvert - \left\lvert\{j \in \{1,\dots,n\}: \theta_j \geq Q(\varepsilon)\}\right\rvert, 0\}}{n}\right)$
INEFF	$\frac{\left\lvert\{j \in \{1,\dots,n\}: \theta_j \leq Q(1-\varepsilon) \cap \hat{\theta}_j \leq Q(1-\varepsilon)\}\right\rvert}{\left\lvert\{j \in \{1,\dots,n\}: \theta_j \leq Q(1-\varepsilon)\}\right\rvert} \cdot \left(1 - \frac{\max\{\left\lvert\{j \in \{1,\dots,n\}: \hat{\theta}_j \leq Q(1-\varepsilon)\}\right\rvert - \left\lvert\{j \in \{1,\dots,n\}: \theta_j \leq Q(1-\varepsilon)\}\right\rvert, 0\}}{n}\right)$
CORRI	$\sum_{k=1}^{\gamma} \frac{1}{\gamma} \frac{\left\lvert\{j \in \{1,\dots,n\} \lvert \lvert\theta_j - \hat{\theta}_j\rvert \leq k \cdot \delta\}\right\rvert}{n}$

Table 2 - Performance indicators used for evaluation

Next, we turn to the quantification of purpose 1) of a DEA and therefore the ability of a model to identify efficient (EFF) and inefficient (INEFF) DMUs. In general DEA, a DMU needs to achieve a score of one to be deemed efficient. As the true efficiency values are drawn from continuous distributions, the probability to draw an exact value of one is zero. However, there exists a positive probability to draw a value of one, because we round to two decimals. Thus, a small interval of different values around one actually result in a drawn value of 1.00. Although, the probability to draw such a value is not zero, it remains very small. Especially in small scenarios, it is possible that no DMU with a value of one is generated, making the identification of such units impracticable. Still, we want to assess if a DEA model is able to identify the top performing units of a sample. Therefore, we define a DMU efficient if its true efficiency score is at least as high as a certain quantile $Q(\varepsilon)$ of the true efficiency distribution. Correspondingly, inefficient DMUs have a true value smaller than or equal to $Q(1-\varepsilon)$. With this flexible (in)efficiency definition, we are as well able to handle different efficiency distributions in the DGP and make the results of various scenarios comparable. Parameter ε should be set high enough, to act as an

acceptable boundary for efficient units. On the other hand, only if a sufficient number of units is deemed efficient, a meaningful analysis on the assessment capability of a model is possible. With regard to the balance between these two goals, we set $\varepsilon = 0.85$ for our study. As MAE only provides an average value without information on the deviation, we track the capability of the models to place their estimates in certain corridors around the true values. The mean value over those corridors results in the indicator CORRI. The tightness of the corridors is set by the parameter δ. In our study, we use a corridor of $\delta = 0.05$. The implication is to test, how good a model is in placing the efficiency estimate at most 5% points apart from its correct value. As a model that has a deviation of 5.1% points in every estimate would fail completely at the indicator, although its results are good, we implemented the possibility to regard more than one corridor. Therefore, parameter γ sets the number of corridors and δ is multiplied with the corridor number $k = 1, \dots, \gamma$. In doing so, we generate corridors of the same size. Due to preliminary testing, we use three corridors in our study, i.e. $\gamma = 3$. Consequently, the corridor boundaries are at 5, 10, and 15% points. As excellent estimates score in all γ corridors, they are implicitly weighted higher. Together with the MAE indicator, CORRI represents purpose 3) of a DEA, where the setting of improvement targets is implicitly achieved by a convincing efficiency estimation.

After generating the data of a scenario and estimating the efficiency scores, we calculate the performance indicators for every model. These indicator values (e.g., MAE) are the average values over all scenarios. The final benchmark value of a model consolidates all performance indicators. On our minds, a good DEA model should be able to perform convincingly over all indicators. For this reason, we define our aggregated indicator B.-Value in Eq. (1).

$$\text{B.}-\text{Value} = \frac{\text{MAE}+\text{SPEAR}+\text{EFF}+\text{INEFF}+\text{CORRI}}{5} \tag{1}$$

However, despite one model always performs slightly better than another one, the B.-Values of these two models can be quite similar. To catch the effect of dominance, we introduce a second aggregated indicator B.-Rank in Eq. (2).

$$\text{B.}-\text{Rank} = \frac{\text{rank(MAE)} + \text{rank(SPEAR)} + \text{rank(EFF)} + \text{rank(INEFF)} + \text{rank(CORRI)}}{5} \tag{2}$$

In a comparison of five models, the model with the best indicator value (e.g., MAE) receives rank one, the worst rank five. Repeating this procedure for every single run of every scenario yields in the end the average indicator rank, e.g. rank(MAE). The B.-Rank is the average over all indicator ranks and depending on the number of models under assessment. Compared to the B.-Value, we do not receive information

comparable to other studies. However, the B.-Rank reveals some interesting insights in the comparison of two specific models.

Both benchmarking scores are stochastic due to the randomly generated data. To ensure robustness of the results, it is essential to conduct a sufficient number of replications for each scenario. A fact, unfortunately neglected by many of the existing Monte Carlo analyses of DEA. The necessary number of replications to ensure sufficient robustness is strongly depending on the scenario. For these reasons, a static replication quantity is not suitable. So far, Zhang & Bartels (1998) are the only ones we are aware of, who are generating their replications flexibly, yielding in 5 up to 100 replications per scenario. For our benchmarking procedure, we created a process in which we start with the conduction of 50 replications for each scenario. After every replication, the B.-Value is recalculated as the average over all replications conducted of the scenario, so far. Afterwards, the moving standard deviation of the B.-Value for the last 25 replications is observed. If this value falls short of 0.1% for all models at the same time, we stop performing further replications. Using only the B.-Value for the stopping criterion is reasonable, as contrary to the B.-Score, since it is independent from the number of models under evaluation. Figure 2 is depicting an example for this procedure.

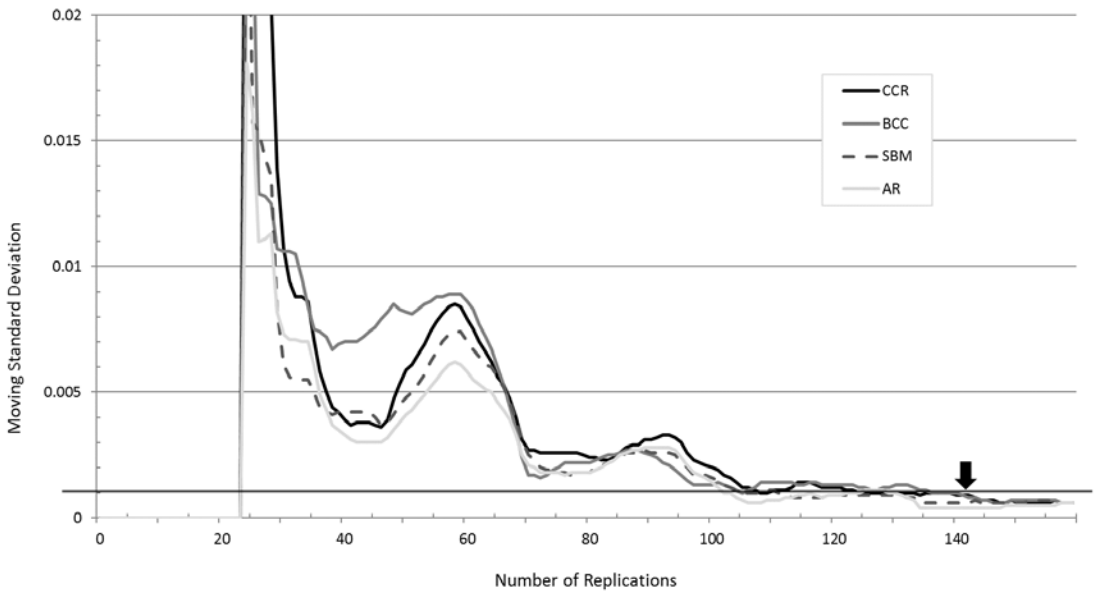


Figure 2 - Example for the moving standard deviation of the B.-Value

Each line is presenting the moving standard deviation of the B.-Value for one DEA model. The calculation starts after 25 replications. The deviation between the B.-Value after 25, 26, 27… replications is decreasing, although results of outlier replications can temporarily raise the standard deviation. In the example, the moving standard deviation of the B.-Value falls short of 0.001 for all models after 143 replications. At this point, we stop performing further replications. The final B.-Value of the scenario is

the average value over all 143 replications. Using this dynamic stopping criterion, we ensure the reproducibility of our study results.

3 Data generation process for one DMU

As in real data sets the true efficiencies are not inherent, these cannot be used for the benchmarking procedure. An artificial data set is necessary for the comparison of different DEA models regarding their efficiency estimation quality. A sophisticated DGP is therefore important to image real data as closely as possible. It stands to reason, that not only one "real" production scenario is existing as companies in different economic sectors face different environmental circumstances. For this reason, the creation of a variety of scenarios is inevitable. We will discuss the settings for the creation of scenarios used in our study in Section 4. For the creation of a meaningful DGP, it is important to keep the need for scenario variations in mind. The goal of the DGP is to create a reasonable single output (y_j) out of meaningful inputs $(x_{ij}, i = 1, \dots, m)$, and true efficiency values (θ_j) for a DMU j. For the sake of simplicity, we drop the DMU index j in the following, as we always talk about a single DMU. For the calculation of the raw output $\tilde{y}$ of a DMU, we utilize the Translog production function given in Eq. (3a), which has been introduced by Christensen et al. (1971, 1973) and has advanced as the gold standard for Monte Carlo simulated data in the past years.

$$\tilde{y} = \prod_{i=1}^{m} x_i^{\alpha_i} \prod_{i=1}^{m} x_i^{\frac{1}{2}\sum_{h=1}^{m} \beta_{ih} \ln x_h} \tag{3a}$$

We use this form for the actual creation of the raw output. For the explanation of characteristics, the logarithmic form of the Translog (3b) is easier to handle and will therefore be used most of the time hereafter.

$$\ln \tilde{y} = \sum_{i=1}^{m} \alpha_i \ln x_i + \frac{1}{2} \sum_{i=1}^{m} \sum_{h=1}^{m} \beta_{ih} \ln x_i \ln x_h \tag{3b}$$

The parameter α_i is predominantly responsible to set the importance of an input i for the production process. The parameter β_{ih} can be used to adjust the substitution properties of the production process between the inputs i and h. The seven steps of the DGP for one DMU are as follows:

1. Set the **number of inputs** (m) for the current DGP,
2. Draw m raw input values from a uniformly distributed **input range,**
3. Adjust the raw inputs with the Corlesky transformation to receive the desired **input correlation,**

4. Choose the **importance of inputs** to set $\alpha_i \; \forall i$ and the input **substitutability** and **substitution distribution** to set $\beta_{ih} \; \forall i, h$,
5. Calculate the single raw output ($\tilde{y}$) by using (3a),
6. Draw a true efficiency value θ for the DMU from the **true efficiency distribution**, and
7. Calculate the final output y by multiplication of the raw output with the true efficiency value $y = \theta \cdot \tilde{y}$.

To ensure a proper data generation, several properties need to be inherent in each production function. According to Coelli et al. (2005), these are:

1) Nonnegativity: The value of the production function is a finite, non-negative, real number,
2) Weak essentiality: The production of positive output is impossible without the use of at least one input,
3) Monotonicity in inputs: Additional units of an input will not decrease the output (also often called free disposability of inputs), and
4) Concavity in inputs: Any linear combination of the vectors x^0 and x^1 will produce an output that is no less than the same linear combination of $f(x^0)$ and $f(x^1)$, i.e., the law of diminishing marginal productivity has to hold.

Throughout the whole DGP, we ensure the adherence to these four properties. They are of special importance for the calibration of the parameters α and β. Apart from α and β, θ and x need to be specified. The phrases in the DGP highlighted in bold are adjustable characteristics that can be used to do so. Their calibration, which will be discussed in the following, aims towards the creation of realistic data.

Generation of $\boldsymbol{\theta}$. The true efficiency value θ is drawn from a random distribution. We assume that DMUs with an extremely low efficiency will not be able to survive and are not present in the market. Furthermore, most studies with sufficient data support show that DMUs with 100% technical efficiency are present, but not predominant on the market. Therefore, a true efficiency value of 1 should not have the highest density in the distribution. We created a true efficiency distribution, where values between a lower bound and 1 can be attained and the mode of the distribution is below 1. We utilize a truncated normal distribution, which we cut at the lower bound and at an upper bound of 1. Different lower bounds can be set to mimic different economies. An according adjustment of the mode and standard deviation keeps the shaping of the distribution similar.

Generation of $\boldsymbol{x}$. The settings on the number of inputs, input range, and input correlation all aim towards the generation of the input vector x. These settings are mainly straightforward as, e.g. changing the number

of inputs. The input range defines the lower and upper bound for the (discrete) uniform distribution we draw the raw inputs from. A small range implies a very homogenous data set with DMUs of rather similar size. A large range on the other hand stands for a more heterogeneous production environment. To assume a correlation between input values is rational. A considerably larger DMU will usually use more inputs than a smaller one. This fact is considered by applying correlation via a Cholesky decomposition (see Hazewinkel 1995) to the inputs. We discuss the values used in our experimental study in Section 4.

Before discussing the next characteristic, we want to recall the purpose of the paper, which is the evaluation of the accuracy of DEA models in CRS settings. Ensuring that the generated data inherit the CRS assumption is therefore indispensable. The creation of meaningful CRS data from a Translog production function is one of the contributions of the paper. Therefore, we have a closer look at the scale elasticity of the Translog production function. The necessary condition for CRS $\phi(x) \stackrel{!}{=} 1$, can be derived from (3b) (see Eq. (4)).

$$\phi(x) = \sum_{i=1}^{m} \frac{\partial \ln \tilde{y}}{\partial \ln x_i} = \sum_i \alpha_i + \sum_i \sum_h \beta_{ih} \ln x_h \stackrel{!}{=} 1 \tag{4}$$

For the realization of global CRS, $\phi(x)$ needs to be equal to one at all possible values of the vector x. Setting all β_{ih} to zero and the $\sum_i \alpha_i$ equal to one obviously achieves this goal. The result is a classical Cobb-Douglas function with the drawback that altering substitution effects is not possible. By considering the symmetry assumption $\beta_{ih} = \beta_{hi}$ (Boisvert 1982, Perelman & Santín 2009) it is possible to re-write Eq. (4) (see Eq. (5)).

$$\phi(x) = \sum_i \alpha_i + \sum_h \left(\beta_{hh} + \sum_{i \neq h} \beta_{ih} \right) \ln x_h \tag{5}$$

From Eq. (5) we can see that sufficient conditions for global CRS that still allow the implementation of substitution effects can be given in Eq. (6).

$$\beta_{hh} = -\sum_{i \neq h} \beta_{ih} \quad \forall h \quad \cap \quad \sum_i \alpha_i = 1 \tag{6}$$

Generation of $\boldsymbol{\alpha}$. Keeping these conditions in mind, we turn back to the setup of characteristics. The implementation of different settings for the importance of the inputs is possible by the choice of α. We

apply two different settings where $\sum_i \alpha_i = 1$ always has to hold to ensure the implementation of CRS. In the first setting (SYM), all inputs are equally important for the production which is given in Eq. (7).

$$\alpha_i = \frac{1}{m} \quad \forall i \tag{7}$$

It can be easily seen that Eq. (7) fulfills the condition $\sum_i \alpha_i = 1$ by Eq. (8):

$$\sum_{i=1}^{m} \alpha_i = \sum_{i=1}^{m} \frac{1}{m} = m \cdot \frac{1}{m} = 1 \tag{8}$$

For the second setting (ASYM), we generated a pattern, where all inputs have different, yet equidistant importance (see Eq. (9)). Input 1 is always the one with the lowest impact on the production. Furthermore, the importance of the inputs is rising with their index. As we only consider abstract inputs and the inputs could be resorted, this regular assignment inherits no distortion of results.

$$\alpha_i = \frac{i + m}{0.5m^2 + 0.5m + m^2} \quad \forall i \tag{9}$$

Please note that Eq. (9) fulfills $\sum_i \alpha_i = 1$ by Eq. (10):

$$\sum_{i=1}^{m} \alpha_i = \sum_{i=1}^{m} \frac{i + m}{0.5m^2 + 0.5m + m^2} = \sum_{i=1}^{m} \frac{i}{0.5m^2 + 0.5m + m^2} + \sum_{i=1}^{m} \frac{m}{0.5m^2 + 0.5m + m^2} = \tag{10}$$

$$\frac{\frac{1}{2}m \cdot (m+1)}{0.5m^2 + 0.5m + m^2} + \frac{m \cdot m}{0.5m^2 + 0.5m + m^2} = \frac{0.5m^2 + 0.5m + m^2}{0.5m^2 + 0.5m + m^2} = 1$$

For pointing out the difference between the SYM Eq. (7) and ASYM Eq. (9) setting, an example with three inputs is chosen. We obtain the values for α_i given in Table 3.

i	1	2	3
SYM	0.333	0.333	0.333
ASYM	0.267	0.333	0.400

Table 3 - Values of α_i for three inputs in SYM and ASYM cases

The effects of the two different α_i settings are displayed in Figure 3. To demonstrate the influence of the SYM/ASYM setting, we show the impact of a change in one of the three inputs on the resulting output. The corresponding input is increased up to 1,000 whereas the other two inputs stay constantly at a value of 50. Performing this approach three times, once for each of the inputs x_1, x_2, and x_3, results in the three lines of Figure 3. For the SYM setting, whatever input is changed, the effects are always the same. For this reason, the three lines lie on top of each other. In the ASYM setting, the importance of every input is different. As input one has the lowest importance for the production process, a rise in $x_1$results in a more moderate rise of the output than a rise of x_2 does. Raising x_3 yields the largest output growth.

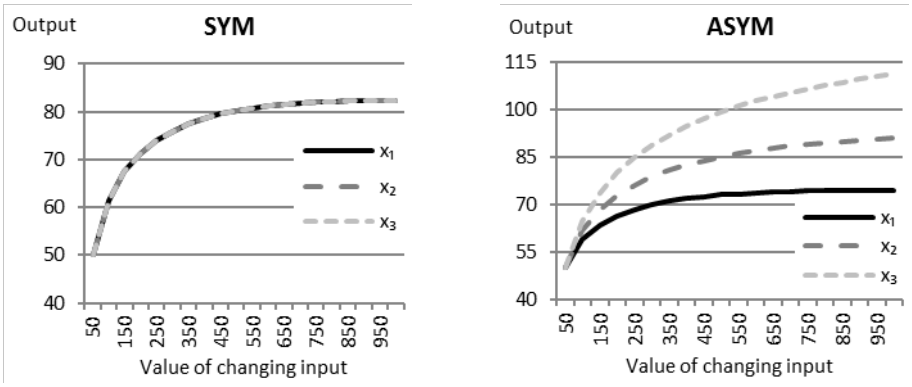


Figure 3 - Example for impact of asymmetry settings; All inputs except for the changing at 50

Generation of $\boldsymbol{\beta}$. Next, we are turning to the substitution of inputs, which can be altered by β in Eq. (3b). Limitations in the setting of the β parameters arise from conditions 3) (Monotonicity of inputs) and 4) (Concavity in inputs) by Coelli et al. (2005). We address these conditions soon. Furthermore, β needs to be symmetric (Boisvert 1982, Perelman & Santín 2009) as given in Eq. (11).

$$\beta_{ih} \stackrel{!}{=} \beta_{hi} \tag{11}$$

Within the boundaries of these conditions, the β values can be set freely. We see two characteristics, which are worth considering in the substitution context. To adjust both characteristics separately, we assume β can be decomposed into two terms: *substitution distribution* and *substitutability* (see Eq. (12)).

$$\beta_{ih} = \overset{\text{substitution distribution}}{\widehat{\sigma_{ih}}} \cdot \overset{\text{substitutability}}{\tilde{\nu}} \tag{12}$$

The characteristic of the substitution distribution accounts for the fact, that substitution might, but does not have to be equal between all inputs. Responsible for the substitution distribution is the distribution of β.

Furthermore, we want to consider varying abilities to substitute inputs. We call this characteristic substitutability and determine its scale by the magnitude of β.

We denote the substitution distribution term by σ_{ih}. As the final magnitude of β_{ih} should be determined by the parameter ν, the σ_{ih} values are normalized between -1 and 1. The implication from Eq. (11) for the substitution distribution term is that $\sigma_{ih} = \sigma_{hi}$. We implement two different settings for the substitution distribution: In the first setting, the substitution between all inputs is equal. In the second setting, the substitution is unequal. For both, equal and unequal substitution distribution settings, Eq. (13) needs to hold to ensure the adherence to Eq. (6) and therefore the implementation of global CRS.

$$\sigma_{hh} = -\sum_{i \neq h} \sigma_{ih} \quad \forall h \tag{13}$$

The imposition of an equal substitution distribution is straightforward and can be achieved by Eq. (14).

$$\sigma_{hh} = -1, \quad \sigma_{ih} = \frac{1}{m-1} \quad \forall i, h, i \neq h. \tag{14}$$

For the realization of unequal substitution scenarios, we use a pattern to generate symmetric but unequal values for σ by Eqs. (15) and (16). A detailed derivation of Eq. (15) and Eq. (16) can be found in the Appendix.

$$\sigma_{ih} = -\frac{m \cdot \left(1.5 - \frac{h-1}{m-1}\right) - \left(2 - 2 \cdot \frac{h-1}{m-1}\right)}{1.5 \cdot m - 2} \quad \forall i, h, i = h \tag{15}$$

$$\sigma_{ih} = \frac{2 - \frac{h-1}{m-1} - \frac{i-1}{m-1}}{1.5 \cdot m - 2} \quad \forall i, h, i \neq h \tag{16}$$

For both cases, Table 4 is showing an example using three inputs. In the equal substitution distribution case, all values on the main diagonal have a value of -1, whereas all others have a value of 0.5. In the unequal substitution distribution case, the differing values lead to a different substitution between inputs. The lower the value, the better the substitution between two inputs, i.e. it is easier to substitute x_2 and x_3 than x_1 and x_2.

Equal substitution distribution				Unequal substitution distribution			
σ_{ih}	$h=1$	$h=2$	$h=3$	σ_{ih}	$h=1$	$h=2$	$h=3$
$i=1$	-1	0.5	0.5	$i=1$	-1	0.6	0.4
$i=2$	0.5	-1	0.5	$i=2$	0.6	-0.8	0.2
$i=3$	0.5	0.5	-1	$i=3$	0.4	0.2	-0.6

Table 4 - Values of σ_{ih} for three inputs in equal and unequal substitution distribution cases

Finally, we turn to the term ν that allows the adjustment of the substitutability of inputs. The necessary monotonicity condition (17) and curvature condition (18) of the production function are playing a key part for the characteristic of regularly behaved production data (Perelman & Santín 2009, Cordero et al. 2015). Remember the points 3) (Monotonicity in inputs) and 4) (Concavity of inputs) of Coelli et al. (2005) discussed above.

$$\frac{\partial \tilde{y}}{\partial x_i} = \frac{\tilde{y}}{x_i}\left(\alpha_i + \sum_{h=1}^{m} \beta_{ih} \ln x_h\right) \overset{!}{\geq} 0 \qquad \forall i \tag{17}$$

$$\frac{\partial^2 \tilde{y}}{\partial^2 x_i} = \frac{\tilde{y}}{x_i^2}\left[\beta_{ii} + \left(\alpha_i + \sum_{h} \beta_{ih} \ln x_h\right)^2 - \left(\alpha_i + \sum_{h} \beta_{ih} \ln x_h\right)\right] \overset{!}{\leq} 0 \qquad \forall i \tag{18}$$

A meaningful DGP needs to ensure that despite of changing substitutability of inputs, a rise in inputs never results in an output decline. This reflects the concept of free disposability of inputs, which is inherent in the vast majority of DEA models. For the evaluation of congestion models, our procedure could be adjusted at this point. Furthermore, to respect the law of diminishing marginal productivity, concavity needs to be assured. For the adherence to the monotonicity and curvature constraints, the magnitude of β plays a crucial role. Therefore, we use the mathematical program (19) (19a)to derive ν. The minimum value for ν implies a "flat" substitution curve and therefore a high substitutability between inputs, whereas the maximum value for ν results in low substitutability.

$$\min / \max \nu \tag{19a}$$

$$\alpha_i + \sum_{h=1}^{m} \sigma_{ih} \cdot \nu \cdot \ln x_h \geq 0 \qquad \forall i \tag{19b}$$

$$\sigma_{ii} \cdot \nu + \left(\alpha_i + \sum_h \sigma_{ih} \cdot \nu \cdot \ln x_h\right)^2 - \left(\alpha_i + \sum_h \sigma_{ih} \cdot \nu \cdot \ln x_h\right) \leq 0 \qquad \forall i \qquad (19c)$$

The example in Figure 4 shows the impact of the different settings on the production process. We consider a simple case of two inputs x_1 and x_2 producing one output. x_1 is increased from 100 to 1,100, while x_2 is decreased simultaneously from 1,100 to 100. Both inputs are equally important for the production, i.e. a SYM setting is chosen. As a result, the highest output is reached for $x_1 = x_2 (= 600)$. From this point on, substituting one input into the other leads to an output reduction. In the case of high substitutability, this reduction is significantly smaller (solid line) than in the case of low substitutability (dashed line).

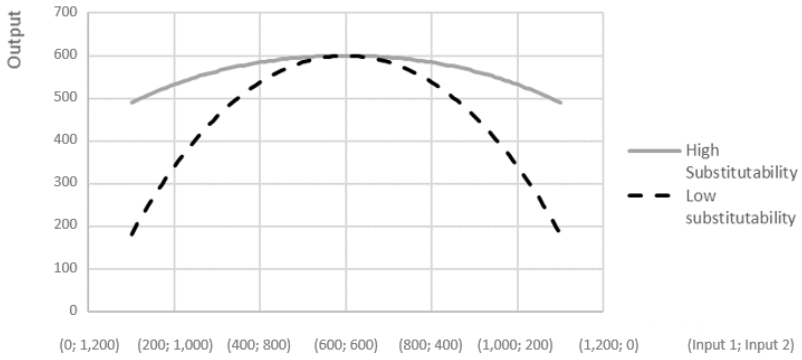


Figure 4 - Substitutability in a 2-input case, where $x_2 = 1200 - x_1$

With the two characteristics on substitution, the DGP for a single DMU is complete. The upcoming section discusses the specific values chosen in each characteristic. Furthermore, it is pointing out the progression from the DGP of a single DMU to the complete experimental study.

4 Study design

Apart from a sophisticated DGP for one DMU, the consideration of a multitude of scenarios is vital for the general validity of results. A scenario is always a concrete combination of values for all characteristics. Apart from the seven characteristics discussed with the DGP for a single DMU, the number of DMUs (n) completes the set of characteristics. So far, Pedraja-Chaparro et al. (1999) created the most scenarios in a Monte Carlo DEA evaluation. As denoted in Table 1, they used 190 different scenarios, generated from three characteristics. For our benchmarking procedure, we generate 1,296 different CRS scenarios, emerging out of all level combinations of our eight characteristics in a multifactorial test design. Compared

to the existing literature, our study design achieves a higher level of validity. Table 5 below is summarizing all eight characteristics with their value specification.

Characteristic	Levels
1. # DMUs (n)	50, 150, 450
2. True efficiencies (θ)	low, medium, high
3. # Inputs (m)	3, 5, 7
4. Input correlation	0, 0.4, 0.8
5. Input range	[100, 1100], [100, 10100]
6. Importance of inputs (β_i)	SYM, ASYM
7. Input substitution distribution (σ_{ih})	equal, unequal
8. Input substitutability (ν)	low, high

Table 5 - Scenarios of the data generation process based on eight distinct characteristics

In the following, we describe these specifications for all characteristics. With the levels for n, we cover an average small (50 DMUs), medium (150 DMUs), and large (450 DMUs) setting for DEA studies. Adjusting the number of DMUs is straightforward, i.e. the DGP for one DMU is repeated n times. The true efficiency values θ are drawn from a random distribution and multiplied with the raw output. In order to create reasonable true efficiency values, the values are drawn from a truncated normal distribution. To test, if the true efficiencies have an influence on our results, we incorporate different true efficiency distributions as characteristics. For the upper efficiency values, truncation is always conducted at 1. The lower bound of the true efficiencies depends on the setting: low, medium, and high. We use lower bounds of 0.25 (low), 0.40 (medium), and 0.55 (high). This restriction of the true efficiency reflects the fact, that in the real world, organizations with extremely low efficiency would not be able to survive. The three settings differ in modes and standard deviations to obtain similar curvatures. We apply modes of 0.75 (low), 0.80 (medium), and 0.85 (high) as well as standard deviations of 0.27 (low), 0.25 (medium), and 0.23 (high). Figure 5 is depicting the probability distributions for the three levels of efficiency.

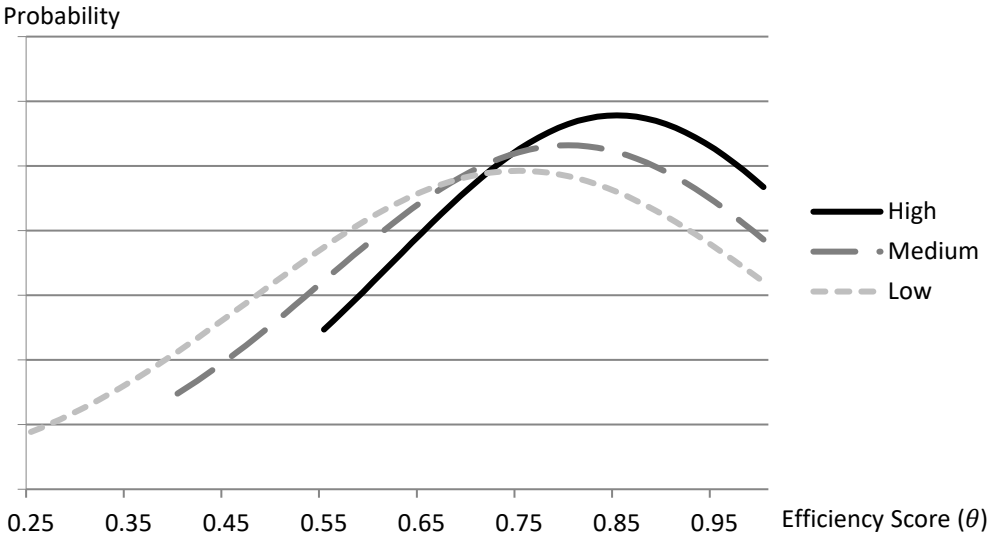


Figure 5 - Distribution of true efficiency scores

The adjustment of the number of inputs (m) is straightforward. For the selection of values for m, we analyzed the number of inputs used in over 260 recently published articles reviewed by Kohl et al. (2018). Based on this experience we set the value for a small study to three, for a medium sized to five, and for a large study to seven. In our DGP, the m inputs for each DMU are drawn from the uniform distributions U[100; 1,100] and U[100; 10,100]. We decided to use these ranges on the basis of a pre-study, where we compared various ranges. Details of the pre-study are discussed in the Appendix. To obtain correlated values, the m raw inputs are transformed by using the Cholesky decomposition (see Hazewinkel 1995) with correlation factors of 0, 0.4, and 0.8. The remaining three characteristics (importance of inputs and importance of substitution, i.e. input substitution distribution and input substitutability) are set according to the results given in Section 3.

5 Evaluated DEA models

After generating an instance of a scenario, the resulting data set is evaluated with four different DEA models (CCR, BCC, AR, and SBM). This is the widest comparison of DEA models with Monte Carlo simulated data so far. In addition, we calculate the benchmark RAND, consisting of randomly drawn numbers from the true efficiency distribution. With RAND, we are able to provide a lower bound for our benchmark scores and facilitate the classification of the B.-Values of other models. The models under evaluation are presented in the following. The model being under high scrutiny is the basic CCR model (Charnes et al. 1978). As in the DGP the true efficiency value is multiplied with the output, an output

orientation is used for all models to receive appropriate results. We present the envelopment form of the CCR model in (20a)-(20d).

$$\max \hat{\theta}_o \tag{20a}$$

$$x_{io} \geq \sum_{j=1}^{n} \lambda_j x_{ij} \qquad \forall i \tag{20b}$$

$$\hat{\theta}_o y_{ro} \leq \sum_{j=1}^{n} \lambda_j \, y_{rj} \qquad \forall r \tag{20c}$$

$$\lambda_j \geq 0 \qquad \forall j \tag{20d}$$

In general, we use a classic notation, where x_{ij} is the amount of input $i = 1, \dots, m$ of DMU $j = 1, \dots, n$, y_{rj} is the amount of output $r = 1, \dots, s$ of DMU $j = 1, \dots, n$. Variable λ_j is the decision variable of the envelopment form and $\hat{\theta}_o$ the resulting estimated efficiency score of the model for the observed DMU o. Although not suitable for the evaluation of CRS data, we also assess the BCC model (Banker et al. 1984), which is the basic model for variable returns to scale (VRS). This assessment can be used as a benchmark and helps to understand the concept of our research. To receive the BCC model, the mathematical program (20) is extended by constraint (21).

$$\sum_j \lambda_j = 1 \tag{21}$$

As both basic models have been discussed excessively in the literature, we refer the interested reader to standard work as Cooper et al. (2007) for further information. The third model we are evaluating is an AR (Assurance Region) model. Assurance region models are DEA models that restrict the values the weights of the inputs and outputs can obtain and were developed by Dyson & Thanassoulis (1988). These weight restrictions prohibit the expelling of inputs or outputs from the analysis by assigning weights of zero. Meanwhile, several ways to restrict the weights of DEA models have emerged. Overviews are provided in Allen et al. (1997) and Pedraja-Chaparro et al. (1997). They show that two dimensions for the restriction of weights in AR models exist: First, the weights to be constrained (raw vs virtual weight restriction). Second, the limits placed on the weights (absolute vs relative weight restrictions). While raw weight restrictions just limit the weight in the primal multiplier model itself, virtual weight restrictions limit the product of weight and input, i.e. $v_i x_{ij}$. Furthermore, absolute restrictions only affect one weight, while relative restriction affect the ratio between two weights. Following Pedraja-Chaparro et al. (1997) we focus on relative weight restrictions. As we, unlike Pedraja-Chaparro et al. (1997), are working with different input elasticities, we chose to apply virtual weight restrictions resulting in the AR model (22). Decision variables v_i and u_r denote as usual the weights for input i and output r. The parameter k for limiting the weights is set in the same way as in Pedraja-Chaparro et al. (1997) as $k = 2$.

$$\min \sum_{r=1}^{s} v_i x_{io} \tag{22a}$$

$$\sum_{i=1}^{m} u_r y_{ro} = 1 \tag{22b}$$

$$\sum_{r=1}^{s} u_r y_{rj} \leq \sum_{i=1}^{m} v_i x_{ij} \qquad \forall j \tag{22c}$$

$$v_i, u_r \geq 0 \qquad \forall i, r \tag{22d}$$

$$\frac{v_i x_{io}}{v_h x_{ho}} \leq k \qquad \forall i, h, i \neq h \tag{22e}$$

The last model we are evaluating is the SBM (Slacks-Based Measure) model, which has been developed by Tone (2001). We use the linear version that is given in (23).

$$\min \tau_o = t - \frac{1}{m} \sum_{i=1}^{m} S_i^- / x_{io} \tag{23a}$$

$$t + \frac{1}{s} \sum_{r=1}^{s} S_r^+ / y_{ro} = 1 \tag{23b}$$

$$t x_{io} = \sum_{j=1}^{n} \Lambda_j x_{ij} + S_i^- \qquad \forall i \tag{23c}$$

$$t y_{ro} = \sum_{j=1}^{n} \Lambda_j y_{rj} - S_r^+ \qquad \forall r \tag{23d}$$

$$t, \Lambda_j, S_i^-, S_r^+ \geq 0 \qquad \forall j, i, r \tag{23e}$$

The model is able to include all slacks (s_i^-, s_r^+) into the efficiency score and thus overcomes one deficiency of many other DEA models. The variable t is only used for computational reasons. It is multiplied with the dual decision variable λ_j, i.e. $\Lambda_j = t \cdot \lambda_j$, and the slacks, i.e. $S_i^- = t \cdot s_i^-$, $S_i^+ = t \cdot s_i^+$ for the sake of linearization. As a development of the additive models, the SBM model has no orientation. Subsequent to the conduction of the efficiency estimation, we assess the quality of the DEA models by using different performance indicators introduced in Section 2.

6 Results

The result section is split into three parts: First of all, we introduce some examples that help to understand the upcoming results. These examples pursue the goal to facilitate the interpretation of the B.-Value. The second part is concerned with the results of our main computational study. Here, we derive in detail, which models show the best performance and look at the driving factors for these results. Finally, we focus on the relevance of the number of DMUs for the result accuracy. Based on our B.-Value, we derive a new rule, how many DMUs have to be present in DEA studies, to ensure a sufficient quality of results.

6.1 Examples

To get a better feeling for the study results and the implications of our B.-Value, we present six small examples which are easy to grasp (see Table 6). Apart from the fact that always nine DMUs are evaluated, the setting of the characteristics for these examples is irrelevant. Rather than looking at the circumstances under which a specific B.-Value occurs, the examples demonstrate the implications that can be drawn from a certain B.-Value. For the sake of simplicity, we only show results for the basic CCR model. For every example, we denote the B.-Value in its headline. Afterwards the five single performance indicators introduced in Section 2, whose mean is the B.-Value, are posed. The bottom part of each example shows key features of the corresponding DEA. This allows a comparison of the true efficiency value and the estimated CCR score as well as the associated ranks for all nine DMUs.

	Example 1: B.-Value = 0.65					**Example 3: B.-Value = 0.75**					**Example 5: B.-Value = 0.85**				
	MAE	SPEAR	EFF	INEFF	CORRI	MAE	SPEAR	EFF	INEFF	CORRI	MAE	SPEAR	EFF	INEFF	CORRI
	0.846	0.593	0.444	1	0.370	0.883	0.814	0.889	0.500	0.667	0.884	0.996	1	1	0.370
DMU	True Efficiency	CCR Score	Absolute Error	True Rank	CCR Rank	True Efficiency	CCR Score	Absolute Error	True Rank	CCR Rank	True Efficiency	CCR Score	Absolute Error	True Rank	CCR Rank
1	0.62	1	0.38	7	1	1	1	0	1	1	0.63	0.74	0.11	7	7
2	0.77	1	0.23	4	1	0.71	0.77	0.06	6	7	0.70	0.81	0.11	6	6
3	0.86	0.96	0.10	2	4	0.50	0.55	0.05	7	8	0.89	1	0.11	1	1
4	0.51	0.60	0.09	9	9	0.80	0.93	0.13	4	4	0.72	0.85	0.13	4	4
5	0.76	0.88	0.12	5	7	0.95	1	0.05	2	1	0.33	0.39	0.06	9	9
6	0.59	0.73	0.14	8	8	0.77	0.85	0.08	5	6	0.71	0.83	0.12	5	5
7	0.75	0.93	0.18	6	5	0.36	0.90	0.54	9	5	0.85	1	0.15	2	1
8	0.94	1	0.06	1	1	0.46	0.49	0.03	8	9	0.74	0.88	0.14	3	3
9	0.81	0.9	0.09	3	6	0.89	1	0.11	3	1	0.54	0.65	0.11	8	8
	Example 2: B.-Value = 0.65					**Example 4: B.-Value = 0.75**					**Example 6: B.-Value = 0.85**				
	MAE	SPEAR	EFF	INEFF	CORRI	MAE	SPEAR	EFF	INEFF	CORRI	MAE	SPEAR	EFF	INEFF	CORRI
	0.920	0.661	0.889	0	0.778	0.914	0.830	0.778	0.667	0.556	0.884	0.962	1	1	0.407
DMU	True Efficiency	CCR Score	Absolute Error	True Rank	CCR Rank	True Efficiency	CCR Score	Absolute Error	True Rank	CCR Rank	True Efficiency	CCR Score	Absolute Error	True Rank	CCR Rank
1	0.99	1	0.01	1	1	0.77	0.93	0.16	7	5	0.80	0.91	0.11	3	4
2	0.84	0.89	0.05	4	5	0.98	1	0.02	1	1	0.66	0.79	0.13	6	5
3	0.68	0.71	0.03	7	9	0.64	0.78	0.14	9	9	0.83	1	0.17	2	1
4	0.70	0.72	0.02	6	8	0.92	1	0.08	2	1	0.32	0.39	0.07	9	9
5	0.57	0.82	0.25	9	6	0.83	0.89	0.06	6	6	0.36	0.43	0.07	8	8
6	0.65	0.93	0.28	8	4	0.84	1	0.16	4	1	0.73	0.94	0.21	4	3
7	0.96	1	0.04	3	1	0.84	0.86	0.02	4	7	0.69	0.77	0.08	5	6
8	0.98	1	0.02	2	1	0.77	0.79	0.02	7	8	0.52	0.63	0.11	7	7
9	0.71	0.73	0.02	5	7	0.89	1	0.11	3	1	0.91	1	0.09	1	1

Table 6 - Examples for B.-Values of 0.65, 0.75 and 0.85

Although the examples are only six snapshots, they represent our experience with the B.-Value and provide helpful indications. They show the difference in the composition of similar B.-Values (the average SD of the performance indicators contributing to a B.-Value in our study over all models is 0.1124). Although Example 1 and 2 end up with the same B.-Value, the CCR model reached the best possible INEFF value in Example 1 and the worst in Example 2. For these examples, the MAE and CORRI indicators behave the other way around. Furthermore, Example 2 shows the best CORRI and MAE values of all examples, although it has a relatively low B.-Value. The SPEAR indicator shows noticeable differences along with different B.-Values. While in the examples with a low B.-Value, the CCR model is barely able to reproduce the correct ordering of the efficiency values, the indicator improves significantly with a rising B.-Value.

Generally spoken, higher B.-Values indicate, as intended, a higher level of result quality. Especially the examples with a B.-Value of 0.65 show very different performances regarding the single indicators. Even if the results are in some parts surprisingly good, in aggregation these examples do not fulfill the demands of a reliable efficiency estimate. The examples with a B.-Value of 0.75 are not exceptionally good. However, we deem the results to be of an acceptable level of quality. From our point of view, models receiving a B.-Value of 0.85 (or better) show a very good performance. For these examples, the CORRI indicator is the only one with some room for improvement. All others already show convincing results. As the results of the examples represent the experiences we made in our computational study, we propose a B.-Value of 0.75 as lower threshold for models providing sufficient quality. In order to be classified as a good estimate, a value of 0.85 or higher is necessary.

6.2 Computational study

With the examples in mind, we turn to the results of our study. As described in Section 5, we evaluate five different models on 1,296 different CRS scenarios. The final B.-Values and B.-Ranks reported in Table 7 are the mean values over all scenarios. Furthermore, we denote the minimum and maximum B.-Values, the models achieve in a single scenario, as well as the standard deviation over all scenarios. The RAND model serves as a lower bound for our benchmarks. On average, 86.33 replications were necessary to terminate a scenario. The maximum number of replications for a scenario was 204. Overall, we created 111,879 replications, with every replication being analyzed by all DEA models. Due to the performance of a sufficient number of replications for each scenario, the results are reproducible if the study is repeated.

Model	**B.-Value**	Min	Max	SD	**B.-Rank**
RAND	0.30	0.27	0.35	0.02	4.79
CCR	0.81	0.40	0.96	0.10	2.86
BCC	0.73	0.36	0.92	0.12	3.89
SBM	0.87	0.42	0.99	0.10	1.54
AR	0.87	0.56	0.98	0.08	1.58

Table 7 - Results of CRS scenarios

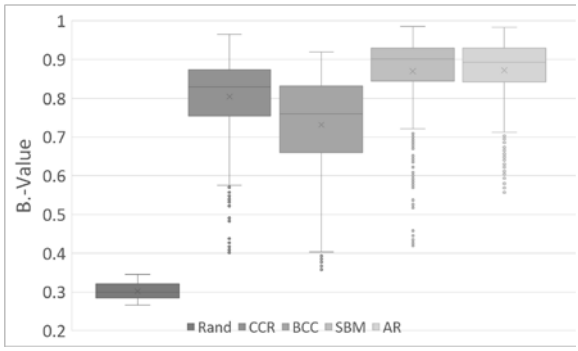


Figure 6 - Boxplots of CRS scenario B.-Value results

The results show several (partly very surprising) observations: First, SBM and AR perform best and considerably better than the CCR model. This is surprising, as the CCR is still the most popular model in applications (Martić et al. 2009). As expected, the BCC model performs worse than the CCR model at the evaluation of CRS data. Yet, this fact is a good indicator for the trustworthiness of the results and helps to understand the mechanics of our method. With a look on the variance of the results, the AR model shows less deviation than SBM, CCR, and BCC. The maximum B.-Values indicate an extremely high quality of

DEA results for some scenarios. The maximum B.-Values of SBM and AR close to 1.0 indicate almost perfect estimates. This becomes even more outstanding if keeping in mind, that these results describe the overall results of a scenario, which is the mean over at least 50 replications. On the other hand, some critical aspects get visible when looking at the minimum B.-Value results. They record weak performances of all models in some scenarios. The AR model is clearly doing best on that score with a minimum value of 0.56. A look at the results in more detail indicates which scenarios are especially concerning. The B.-Rank results, where 1.0 is the best possible value, support in large parts the findings of the B.-Value. However, the SBM model is performing slightly better than the AR model. As explained in Section 2, the B.-Rank depends on the number of analyzed models. It is not only a mere measure of dominance on the average scenario level, but it looks at every indicator in every replication. A B.-Rank value of exactly one is no longer reachable for a model. Assume a single indicator in any replication is better for another model. As a consequence B.-Rank is larger than 1.0. As the B.-Rank is affected by the number of models, we will conduct pairwise comparisons with the B.-Rank for selected models with separate studies.

Analysis of characteristics. In the following, we are addressing the detailed analysis of all 1,296 scenarios. The investigation aims at the identification of patterns and trends which are provoked by the eight different characteristics. Figure 7 provides an overview over the B.-Values of all analyzed CRS models for all 1,296 scenarios.

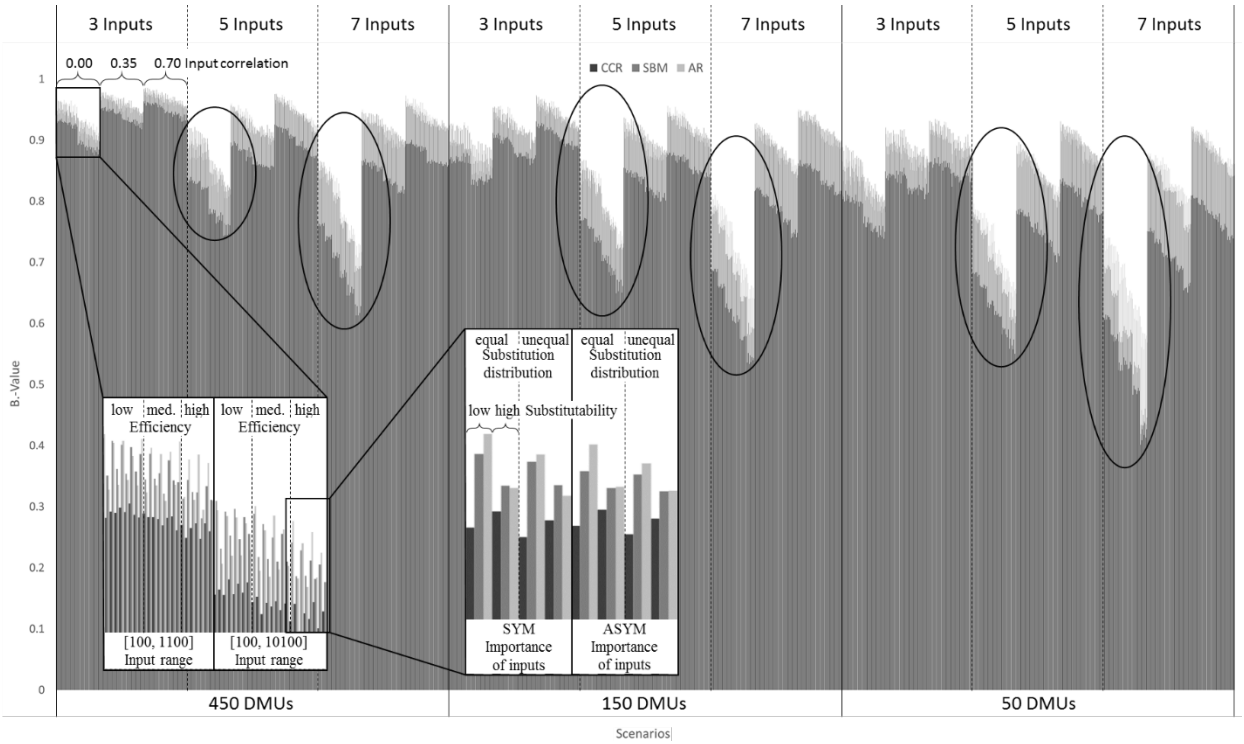


Figure 7 - Results of all 1,296 scenarios for the CRS models

Every bar depicts the B.-Value of a model for a single scenario. The scenarios are sorted starting with characteristic 1 (# DMUs) in a descending order on the highest level. The second level of sortation is characteristic 3 (# inputs), followed by characteristic 4 (input correlation). The further sortation rules can be obtained from the magnified areas of the figure. These sortation rules have been chosen to emphasize as many findings as possible.

These detailed results exhibit clearly the dominance of the SBM and AR model over the CCR model. The darker bars, visualizing the CCR results, stay below the brighter AR and SBM results in every single scenario. Turning to the main drivers of these B.-Value results, several coherences are visible. Recent studies emphasize the large influence of the number of DMUs and inputs on the accuracy of DEA models. Our study can confirm these findings, as shown subsequently. However, these two characteristics are only partly responsible for the most distinctive effect visible in Figure 7. Six large gaps (highlighted with circles) characterize Figure 7. They reveal the huge impact of the correlation of inputs on the results. Missing correlation between the inputs is causing the six gaps. The interdependence between the correlation and the number of inputs is both coherent and visible. The use of more inputs is amplifying the negative effect of a low correlation. Furthermore, the use of a low number of DMUs is as well amplifying the effect. For instance, the average B.-Value for the CCR model drops for 7 inputs, 50 DMUs, and 0.0 correlation to a value of 0.53. Soothing is the high improvement of the estimates already with the inclusion of a moderate correlation of 0.35. A complete absence of correlation between inputs is for most real processes rather unlikely. As already indicated, our study confirms the overall influence of the number of DMUs and inputs on the quality of the DEA estimates. The more units are evaluated, the higher the B.-Value and thus the quality of the results (see aggregates in Figure 8). On a related note, it seems fitting that a setting with more DMUs needs on average less replications to generate stable results. A very similar impact has a change in the number of inputs (see Figure 9). For both effects, our results indicate that AR and SBM are more stable with regard to critical scenarios than the CCR model. In other words, their B.-Value deteriorates less than the CCR's.

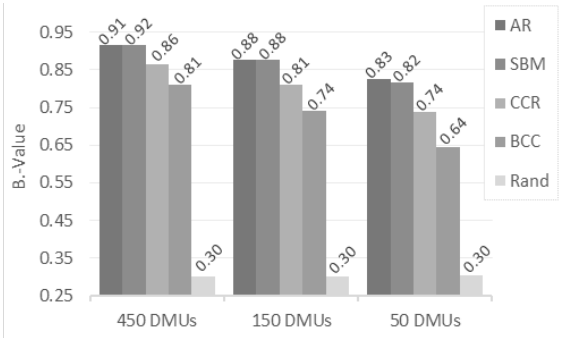


Figure 8 - Impact of a change in DMU size on the B.-Value

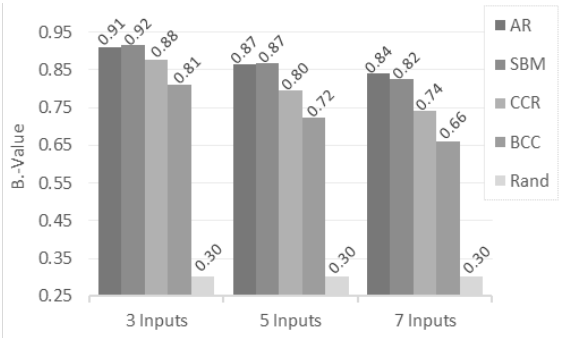


Figure 9 - Impact of a change of the Input number on the B.-Value

The saw tooth appearance of Figure 7 comes from the characteristics 5 (input range) and 2 (true efficiency distribution). It becomes evident, that larger input ranges lead to worse results for all models. A rise in the true efficiency is affecting all models negatively. Interestingly, this effect is fortified by an increase in the number of inputs. In this context, two contradictory effects become apparent. First, a rise in the true efficiency pushes more units close to the true efficiency frontier. This helps all models to reproduce the frontier more precisely. As a result, the MAE and CORRI indicators are rising. Second, a rise in true efficiency squeezes the values to a smaller range. The consequence is a decline in SPEAR, EFF, and INEFF as discrimination becomes more difficult. Overall, the negative effects prevail slightly. However, we see a smaller effect for BCC and AR. These effects are affected by the number of inputs and they grow with an increase in inputs. The effects of input range and true efficiency distribution combined lead to a decline within each group of scenarios having the same #DMUs, #Inputs, and correlation. Variation in the characteristics 6 (importance of inputs), 7 (input substitution distribution), and 8 (input substitutability) have almost no significant effects. Only the AR and SBM model perform marginally worse if the input substitutability is high.

Analysis of indicators. After analyzing the results on basis of the eight different characteristics, we turn to the indicator level. Table 8 depicts the average indicator values over all 1,296 scenarios and their corresponding replications for every model. Remember, the average over all indicator values of a model in turn yields its B.-Value. A support of the previous results by all indicators is visible as AR/SBM $\succ$ CCR $\succ$ BCC $\succ$ RAND is valid for every single indicator.

Indicator	RAND	CCR	BCC	SBM	AR	Avg.
MAE	0.82	0.91	0.88	0.94	0.93	0.897
SPEAR	0.00	0.89	0.79	0.92	0.95	0.710
EFF	0.16	0.76	0.69	0.77	0.83	0.642
INEFF	0.16	0.83	0.76	0.91	0.89	0.711
CORRI	0.36	0.63	0.53	0.81	0.76	0.620
Avg. ($\widehat{=}$ B.−Value)	0.30	0.81	0.73	0.87	0.87	

Table 8 - Performance of the DEA models on the indicator level

While MAE yields on average the best results, CORRI has plainly the lowest value of all indicators. This captures the fact, that DEA estimates are on average quite good, but getting really close to the correct values for the majority of DMUs is nothing that should be expected. The high value of RAND with regard to MAE is comprehensible as the values for RAND are drawn from the same range as the true efficiency values. Therefore, even RAND values that differ as far as possible from the truth do not receive a MAE score of zero. Encouraging is the good performance of most models with regard to the SPEAR indicator. The correct ordering of the DMUs efficiency estimates is essential for the credibility of DEA analyses.

Especially, the outstanding result of the AR model (SPEAR score of 0.95) has to be acknowledged in this regard. Furthermore, a huge difference between the estimates of the DEA models and RAND is apparent. RAND is not (and should not be) able to represent the correct ordering of the DMUs. Concerning EFF and INEFF, all models are better in identifying inefficient units than in identifying the efficient ones. An explanation for this observation is the overestimation of efficiency in large parts by the tested DEA models. Consequently, the unjustified declaration of DMUs as efficient is depressing the EFF indicator.

When comparing the top performing SBM and AR models, it becomes apparent that despite identical B.-Values some differences on the indicator level exist. As mentioned before, the SPEAR value of the AR is exceptionally good and cannot be reached by the SBM. Moreover, the AR model is better in identifying efficient units. In return, SBM has slight advantages in MAE and INEFF and is clearly the best model with regard to CORRI.

Analysis of B.-Ranks. Finally, we are analyzing the B.-Rank results. The comparison of all models with the B.-Rank in Table 7 shows a slight dominance of the SBM over the AR model and a clear dominance of these two over the other models. As the B.-Rank is, unlike the B.-Value, depending on the number of models in the analysis, we conduct three additional studies where only two models are compared (CCR-SBM, CCR-AR, SBM-AR). With this setting, a B.-Rank of 1.0 is the best and a score of 2.0 the worst possible value. A result of 1.0 implies a model is at least as good as its opponent for all indicators in every replication in 100% of the cases. The comparison of the SBM and AR model with the CCR model by the B.-Rank is quantifying their dominance. In a study with the SBM and CCR models only, the SBM model was able to reach a B.-Rank value of 1.05, while the B.-Rank value of the CCR was 1.84. The same comparison between the AR and CCR model in a separate study returned likewise a B.-Rank value of 1.05 for the AR model. The CCR model received a B.-Rank value of 1.90. Based on these surprisingly unambiguous results, we advocate for a supersession of the CCR model as the standard model for the application of DEA. For the comparison of the top performing models, a slight advantage of the SBM (1.43) over the AR (1.48) is visible. These results support the findings of Table 7. Hence, the SBM performs more often at least as good as the AR model. Therefore, we endorse the usage of the SBM model as standard DEA model.

6.3 A guideline for the proper use of the DMU quantity in DEA studies

The importance of a correct setup of DEA studies with regard to the ratio of the number of DMUs to inputs and outputs is well known. The results in Section 6.2 already exhibited increasing study accuracy with a rise in the number of DMUs or a reduction in the number of inputs. Other studies with Monte Carlo simulated data show similar findings for the CCR model (Smith 1997, Pedraja-Chaparro et al. 1999). In our study, we are able to quantify this effect and show its severity. DEA in general tends to overestimate efficiency. The increasing accuracy of studies with a higher number of DMUs results among others from

a reduced overestimation. This becomes apparent, when comparing the overall average true efficiency values with their estimated counterparts for the different levels of characteristic 1 (#DMUs) in Table 9. With more DMUs, the values decrease for all models and get closer to the true score of 0.74.

	True Efficiency	CCR	BCC	SBM	AR
50 DMUs	0.74	0.86	0.90	0.82	0.83
150 DMUs	0.74	0.83	0.86	0.79	0.80
450 DMUs	0.74	0.80	0.83	0.77	0.79

Table 9 - Comparison of average overall results for 50, 150, and 450 DMUs

In this context, a look at existing guidelines for a proper setup of DEA studies with regard to the minimum number of DMUs is worthwhile. A selection of guidelines can be found in Table 10 below. Parameter n denotes the number of DMUs, m gives the number of inputs, and s the number of outputs.

Author	**Minimum number of DMUs**
Golany & Roll (1989)	$n \geq 2(m+s)$
Boussofiane et al. (1991)	$n \geq (m \cdot s)$
Bowlin (1998)	$n \geq 3(m+s)$
Dyson et al. (2001)	$n \geq 2(m \cdot s)$
Cooper et al. (2007)	$n \geq \max\{(m \cdot s); 3(m+s)\}$

Table 10 - Guidelines for minimum number of DMUs

All of these guidelines endorse a study with 7 inputs, 1 output, and 50 DMUs. However, our results show that the quality of such a study is miserable. With this setup, the CCR model reaches a B.-Value of 0.67. With regard to some indicators, it even performs as bad as RAND, i.e. drawing random numbers. To exhibit the relationship between the number of DMUs and the number of inputs, we conduct new calculations. The goal is to reveal for settings with different numbers of inputs, how many DMUs are necessary to receive always the same result accuracy. Therefore, we conduct a study regarding two inputs and examine, how many DMUs are necessary to exceed a predetermined B.-Value with the CCR model. The same procedure is repeated for 3, 4, 5, 6, and 7 inputs. In comparison to our main study in Section 6.2, where 1,296 scenarios have been analyzed, the characteristics 1 (#DMUs) and 3 (#Inputs) are fixed. As both characteristics had three different levels in Section 6.2, the resulting B.-Values for the new studies are the

average values over 144 $\left(=\frac{1{,}296}{3 \cdot 3}\right)$ scenarios. The results for a predetermined B.-Value of 0.75 and 0.80 are depicted in Figure 10.

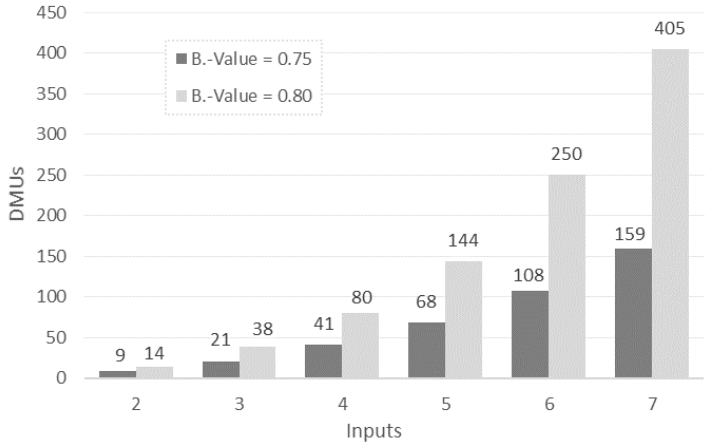


Figure 10 - DMU/Input ratios resulting in constant B.-Values

Three observations become apparent:

1) The number of DMUs needed to ensure a sufficient level of quality in DEA is far higher than suggested by the existing guidelines.
2) There is no linear relationship between the number of inputs and the number of DMUs when considering a constant level of quality. This fact is as well not captured by any of the existing guidelines.
3) The curve with a B.-Value of 0.80 has a far steeper ascend than the curve with a B.-Value of 0.75. To conduct studies with a higher level quality, depending on the number of inputs, a considerable number of additional DMUs might be needed.

Overall, it is important not to be wasteful on the use of additional inputs. A more precise reproduction of the production process might otherwise be displaced by a far less accurate efficiency estimation. To capture observations 1) and 2) above, we propose the following rule for the minimum number of DMUs used in DEA studies in (21).

$$n \geq 20 + \frac{m+s-1}{2} \cdot \left(-10 + 10 \cdot (m+s-3)\right) \tag{24}$$

The new rule presents a reasonable approximation of the CCR curve with a B.-Value of 0.75. Remember, we identified 0.75 as the necessary B.-Value for an acceptable level of quality. The result seems a fitting measure for the restriction of the minimum number of DMUs for DEA in general. For the creation of the

rule, we utilized the property of the results to resemble an arithmetic series. The difference between two consecutive elements follows the pattern $a_{i+1} = a_i + 10$. Tests with eight and nine inputs revealed a good fit of the rule for higher numbers of m as well. Although our DGP includes only a single output, a generalization for inputs and outputs is vital. As in many existing rules inputs and outputs are interchangeable, it stands to reason to include outputs in our rule in the same way.

7 Conclusion

In this paper, we provide a method to assess the quality of DEA model estimations and make their performance visible for everyone interested in DEA applications. We utilize Monte Carlo simulation based on a Translog production function for the underlying data generation process. The generation of a multitude of meaningful scenarios is playing a key role. We are generating 1,296 scenarios with constant returns to scale and evaluate the results by the use of five different performance indicators. Our research enhances for the first time a comparison of the level of quality of DEA models. We can show that the CCR model, which is still state of the art for evaluations with CRS settings, is performing worse than AR and SBM models. As a consequence, we advocate for a rise in AR and SBM applications. Remarkable in this context is the prominent position of the AR model with regard to the absence of a special calibration of the weight restriction. As the SBM model is performing in more scenarios at least as good as the AR model than vice versa, we endorse the establishment of the SBM model as the standard DEA model. Regarding the influence of scenario parameters, we can show that the number of DMUs, the number of inputs, and the correlation between inputs has a major influence on the quality of DEA analyses. While some of these interdependencies were already established, the first time quantification of these effects allows the formulation of a new rule for the minimum number of DMUs. This rule should be used in DEA analyses, to achieve a reasonable level of quality for the estimates. If the adherence to this rule is not possible, e.g. due to a strong limitation in the number of DMUs, a DEA should not be conducted.

Future research could extend our methodology to variable returns to scale scenarios. Another drawback of the analysis is the use of just a single output in the DGP. Perelman & Santín (2009) showed how to extend the Translog DGP to a two input, two output setting. The generalization to a meaningful multi-input, multi-output DGP is however not trivial and leaves room for future research. Furthermore, we want to address the negligence of noise in the data generation process. The reason behind this is the idea to create awareness for the accuracy of DEA results in a perfect setting. We intend to initiate further validation of DEA models with the presented method, both for other already existing models and for new developments. In this way the faith in DEA results could be strengthened and DEA might be able to leave the scientific stage and receive more attention by managers, economists, and politicians.

8 Appendix

8.1 Derivation of unequal substitution distribution

The parameter σ_{ih} needs to be designed symmetric ($\sigma_{ih} = \sigma_{hi}$) and (13) has to hold. Furthermore, the final values should be normalized to the interval $[-1, 1]$. The starting point for the idea of the derivation of unequal substitution distribution was to reduce σ_{ih} by an equidistant amount with a rise in i or h and replace the values on the main diagonal later on to satisfy (13). An example for $m = 3$, which is satisfying our desire for symmetry and equidistant differences before the replacement of the values on the main diagonal would be $\sigma'_{ih} = \begin{pmatrix} 2 & 1.5 & 1 \\ 1.5 & 1 & 0.5 \\ 1 & 0.5 & 0 \end{pmatrix}$. A general formula to reproduce this pattern for m inputs yields in Eq. (25).

$$\sigma'_{ih} = 2 - \frac{i-1}{m-1} - \frac{h-1}{m-1} \quad \forall i, h. \tag{25}$$

In order to satisfy (13), the values on the main diagonal need to be replaced by the sum of the remaining values in their row, multiplied by -1. Since the σ'_{ih}-values build an arithmetic sequence, $\sum_i \sigma'_{ih}$ can be computed by $\frac{m}{2}(\sigma'_{i1} + \sigma'_{im}) = \frac{m}{2} \cdot \left(2 - \frac{i-1}{m-1} - \frac{1-1}{m-1} + 2 - \frac{i-1}{m-1} - \frac{m-1}{m-1}\right) = m \cdot \left(1.5 - \frac{i-1}{m-1}\right)$. In the example above, we would receive for $h = 1, 2, 3$ the values $4.5, 3, 1.5$. The values on the main diagonal have to be the negation of $\sum_{i,i\neq h} \sigma'_{ih} = m \cdot \left(1.5 - \frac{i-1}{m-1}\right) - \left(2 - 2 \cdot \frac{i-1}{m-1}\right)$. We receive Eq. (26).

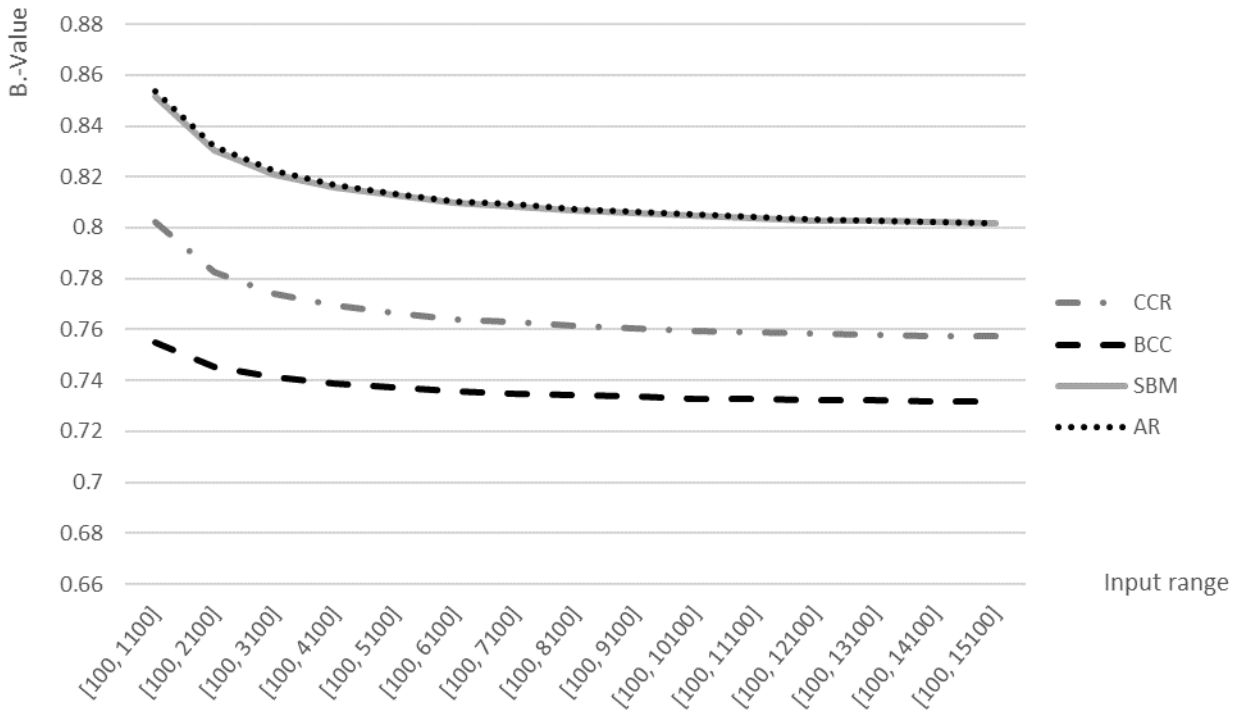
$$\sigma'_{ih} = -m \cdot \left(1.5 - \frac{i-1}{m-1}\right) - \left(2 - 2 \cdot \frac{i-1}{m-1}\right) \quad \forall i = h \tag{26}$$

In the example, $\sigma_{11}, \sigma_{22}, \sigma_{33}$ take the values $-2.5, -2, -1.5$. Finally, in order to obtain values between -1 and 1, all values are divided by the absolute value of the largest element $\left[|\sigma_{11}| = m \cdot \left(1.5 - \frac{1-1}{m-1}\right) - \left(2 - 2 \cdot \frac{1-1}{m-1}\right) = 1.5 \cdot m - 2\right]$. Dividing (25) and (26) by $1.5 \cdot m - 2$ leads to the formulas (15) and (16).

8.2 Pre-study on input ranges

In a pre-study, we tested the effect of changing input ranges. To keep the computational time of this study manageable, only scenarios with 50 DMUs were considered (6,480 scenarios in total). The pre-study shows, that all models react similarly to a change of the input range. An increase of the range is decreasing the B.-Value of all models. After a rather strong B.-Value decrease in the lower input ranges, the values

settle down around an input range of 6,000. To include a high and a low value into our analysis, we adopt input ranges of 1,000 and 10,000 for our DGP.



9 References

Allen, R., Athanassopoulos, A., Dyson, R., & Thanassoulis, E. (1997). Weights restrictions and value judgements in Data Envelopment Analysis: Evolution, development and future directions. *Annals of Operations Research, 73*, 13–34xxx. https://doi.org/10.1023/A:1018968909638

Banker, R. D., Charnes, A., & Cooper, W. (1984). Some Models for Estimating Technical and Scale Inefficiencies in Data Envelopment Analysis. *Management Science, 30*, 1078-1092. Retrieved from http://www.jstor.org/stable/2631725

Boisvert, R. N. (1982). *The Translog Production Function: Its Properties, Its Several Interpretations and Estimation Problems*. Cornell University, NY.

Boussofiane, A., Dyson, Robert, & Thanassoulis, Emmanuel. (1991). Applied data envelopment analysis. *European Journal of Operational Research, 52*, 1–15. https://doi.org/10.1016/0377-2217(91)90331-O

Bowlin, W. F. (1998). Measuring performance: An introduction to data envelopment analysis (DEA). *The Journal of Cost Analysis, 15*, 3–27.

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

Christensen, L. R., Jorgenson, D. W., & Lau, L. J. (1971). Conjugate Duality and the Transcendental Logarithmic Production Function. *Econometrica, 39*, 225–256.

Christensen, L. R., Jorgenson, D. W., & Lau, L. J. (1973). Transcendental logarithmic production frontiers. *The Review of Economics and Statistics, 55*, 28–45.

Coelli, T. J., Rao, D. S. P., O'Donnell, C. J., & Battese, G. E. (2005). *An introduction to efficiency and productivity analysis*: Springer Science & Business Media.

Cooper, William, Seiford, L. M., & Tone, K. (2007). *Data envelopment analysis: a comprehensive text with models, applications, references and DEA-solver software*: Springer Science & Business Media.

Cooper, WilliamW., Ruiz, J. L., & Sirvent, I. (2011). Choices and Uses of DEA Weights. In W. W. Cooper, L. M. Seiford, & J. Zhu (Eds.), *International Series in Operations Research & Management Science. Handbook on Data Envelopment Analysis* (Vol. 164, pp. 93–126). Springer US. https://doi.org/10.1007/978-1-4419-6151-8_4

Cordero, J. M., Santín, D., & Sicilia, G. (2015). Testing the accuracy of DEA estimates under endogeneity through a Monte Carlo simulation. *European Journal of Operational Research, 244*, 511–518.

Dyson, Robert, 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*, 245–259. https://doi.org/10.1016/S0377-2217(00)00149-1

Dyson, Robert, & Thanassoulis, Emmanuel. (1988). Reducing weight flexibility in data envelopment analysis. *Journal of the Operational Research Society*, *39*, 563–576.

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

Golany, B., & Roll, Y. (1989). An application procedure for DEA. *Omega*, *17*, 237–250. https://doi.org/10.1016/0305-0483(89)90029-7

Hazewinkel, M. (Ed.). (1995). *Encyclopaedia of mathematics* (Unabridged reprint of the original 10-vol. hardbound library ed.). Dordrecht: Kluwer.

Holland, D. S., & Lee, S. T. (2002). Impacts of random noise and specification on estimates of capacity derived from data envelopment analysis. *European Journal of Operational Research*, *137*, 10–21.

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*, 137–148.

Martić, M., Novaković, M., & Baggia, A. (2009). Data envelopment analysis-basic models and their Utilization. *Organizacija*, *42*, 37–43.

Pedraja-Chaparro, F., Salinas-Jimenez, J., & Smith, P. (1997). On the role of weight restrictions in data envelopment analysis. *Journal of Productivity Analysis*, *8*, 215–230.

Pedraja-Chaparro, F., Salinas-Jiménez, J., Smith, P., & others. (1999). On the quality of the data envelopment analysis model. *Journal of the Operational Research Society*, *50*, 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*, 303–310.

Siciliani, L. (2006). Estimating Technical Efficiency in the Hospital Sector with Panel Data. *Applied Health Economics and Health Policy*, *5*, 99–116. https://doi.org/10.2165/00148365-200605020-00004

Smith, P. (1997). Model misspecification in Data Envelopment Analysis. *Annals of Operations Research*, *73*, 233–252. https://doi.org/10.1023/A:1018981212364

Tone, K. (2001). A slacks-based measure of efficiency in data envelopment analysis. *European Journal of Operational Research*, *130*, 498–509. https://doi.org/10.1016/S0377-2217(99)00407-5

Yu, C. (1998). The effects of exogenous variables in efficiency measurement—a Monte Carlo study. *European Journal of Operational Research*, *105*, 569–580.

Zhang, Y., & Bartels, R. (1998). The effect of sample size on the mean efficiency in DEA with an application to electricity distribution in Australia, Sweden and New Zealand. *Journal of Productivity Analysis*, *9*, 187–204.

A3. Using Data Envelopment to Estimate Hospital Efficiencies – A Teaching Case

Kohl, S. (2019). Using Data Envelopment to Estimate Hospital Efficiencies – A Teaching Case.

Status: Under revision; submitted June 28, 2019. *MSOR Connections*, no category available.

Using Data Envelopment Analysis to Estimate Hospital Efficiencies – A Teaching Case

Sebastian Kohl[a,b]

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

[b] University Center of Health Sciences at Klinikum Augsburg (UNIKA-T), Neusässer Straße 47, 86156 Augsburg, Germany

* Corresponding author: sebastian.kohl@unikat.uni-augsburg.de, +498125986446

Abstract

This case study has the intention to guide through the conduction of a meaningful Data Envelopment Analysis (DEA) in the healthcare sector. A data sample on German hospitals is provided and used throughout different tasks. Apart from the implementation of the DEA model itself, the case study also covers areas of pre- and post-processing. As a result, the user of the case study is confronted with common pitfalls and learns to work with procedures, which have emerged as gold standards. The participant is encouraged to use methods for the detection of outliers and for the treatment of missing values to cover common issues in this field. The comparison of different DEA models enhances the understanding of the mechanics of DEA, especially the relevance of slacks for the analysis. In including quality data into the study, another essential feature for hospital analyses is addressed.

With the Helmsman DEA, an interesting, however, rather unfamiliar procedure is presented to achieve a meaningful inclusion of the quality data into the analysis. Using bootstrapping as a subsequent method completes the study. Finally, a recommendation for the grading of the tasks is given. The results and the source code to all implementations are provided.

Keywords: Data Envelopment Analysis, Efficiency Estimation, Hospital, Healthcare, Quality, Bootstrapping, Case Study

1 Introduction

The health care sector represents one of the essential parts of social welfare in every country. It is common consent that due to an aging population, the importance and demand for health services in Germany will further increase in the upcoming years. Hospitals are in almost all countries a crucial pillar of the healthcare system. With a decrease in hospital numbers over the past decades and a declining length of stay of patients as a consequence of the introduction and revision of the DRG system, the supply side of the hospital market is shrinking (Klauber et al. 2019). Putting an increasing pressure for hospitals to work at least cost-covering on top of it, a rise in efficiency is without an alternative for hospitals and the healthcare sector. However, the identification of efficient best-practice examples to learn from is not an easy task for hospital managers. As a decision maker, you look at the literature for efficiency estimation and identify two main methodological directions: parametric and nonparametric methods. Parametric approaches as regression and Stochastic Frontier Analysis usually have the disadvantage that a functional form of the production (or service) process needs to be determined. Furthermore, many parametric models only allow for a single output. For these reasons, the nonparametric methods are used more frequently when it comes to the assessment of efficiency in hospitals (Jacobs et al. 2006). Within the non-parametric methods, data envelopment analysis (DEA) models have prevailed as state of the art for multiple inputs, multiple outputs settings. When focusing more on the setup of DEA studies, you learn that for the execution of a meaningful DEA study, the usage of homogenous decision making units (DMUs), in your case hospitals, is of prime importance (Dyson et al. 2001). From your experience as a decision maker, this is reasonable, as small and specialized private hospitals cannot be compared with large university hospitals providing maximum care for an entire region. Therefore, it makes sense for you to limit the analysis to a specific hospital size. The hospital you are working for is a medium-sized hospital. In 2017, the German hospital sector encompasses 1,942 hospitals, with 1,592 hospitals being deemed as general hospitals (Federal Statistical Office of Germany 2018). The average size of a general hospital in Germany is 282.9 beds. Figure 1 shows the distribution of the size of general hospitals in more detail.

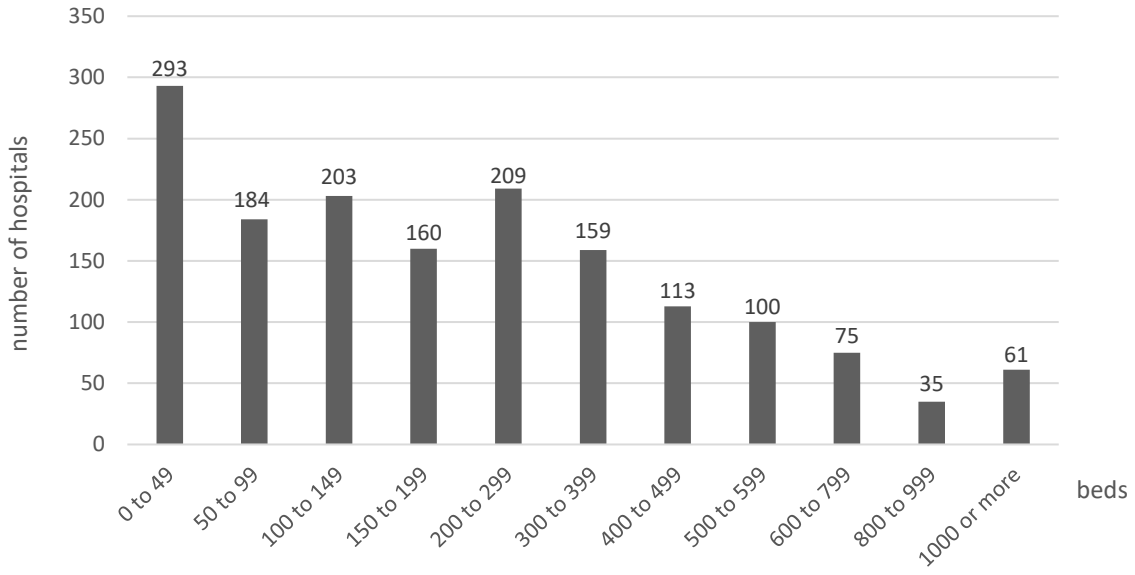


Figure 1: number of general hospitals according to bed clusters (Federal Statistical Office of Germany 2018)

The cluster between 200 and 299 beds seems a fitting choice for your analysis. It inherits your hospital and provides a sufficient amount of DMUs. As the cluster represents as well the average hospital size, the results are also representing large parts of the German hospital Markert in general. A further advantage of this bed range is the automatic exclusion of university hospitals. University hospitals have different objectives, as research and training of residents are additional goals. Therefore, the mixture of hospital types might distort the results.

Restricted by the bed size corridor, you want to compare the performance of the following hospitals (Table 1):

DMU	Hospital	Departments	Ownership	CMI
1	Agaplesion Ev. Bathildiskrankenhaus Bad Pyrmont	7	private non-profit	1.036
2	Asklepios Klinik Lich GmbH	5	private for-profit	1.01
3	Borromäus-Hospital Leehr gGmbH	9	private non-profit	0.885
4	DIAKOMED gGmbH Diakoniekrankenhaus Chemnitzer Land	8	private non-profit	0.955
5	Diakoniekrankenhaus Halle	10	private non-profit	1.226
6	Diakonissenkrankenhaus Dresden	10	private non-profit	0.875
7	Dominikus Krankenhaus GmbH Berlin	4	private non-profit	1.407
8	Donau-Ries-Klinik Donauwörth	7	public	0.829
9	DRK Krankenhaus Luckenwalde	9	private non-profit	0.978
10	Elbe Klinikum Buxtehude	5	public	0.911
11	Ev. Diakonissenkrankenhaus Leipzig	12	private non-profit	1.071
12	Ev. Krankenhaus Ludwigsfelde-Teltow	4	private non-profit	0.87
13	Ev. Krankenhaus Mettmann GmbH	6	private non-profit	0.921
14	Evangelische Krankenhaus Bethanien Iserlohn gGmbH	8	private non-profit	0.793
15	Gemeinschaftsklinikum Mittelrhein, St. Elisabeth Mayen	7	private non-profit	1.024
16	Gesundheitszentrum Tuttlingen	6	public	0.904
17	Heilig Geist Krankenhaus Köln	7	private non-profit	0.876
18	Helios Albert-Schweitzer-Klinik Northeim	11	private for-profit	0.979
19	Helios Klinik Lutherstadt Eisleben	7	private for-profit	0.953
20	Helios Klinik Köthen	5	private for-profit	1.005
21	Helios Klinik Rottweil	11	private for-profit	0.94

22	Helios St. Marienberg Klinik Helmstedt	5	private for-profit	1.084
23	Hospital Zum Heiligen Geist Kempen	7	private for-profit	0.757
24	Josephs-Hospital Warendorf	8	private non-profit	0.925
25	Katholische Kliniken Ruhrhalbinsel (St. Josef Krankenhaus Kupferdreh)	9	private non-profit	1.05
26	St. Marien-Hospital Oberhausen	9	private non-profit	0.958
27	Katholisches Krankenhaus Dortmund-West	4	private non-profit	0.878
28	AMEOS Klinik am Bürgerpark Bremerhaven	5	private for-profit	0.996
29	Klinik Vincentinum Augsburg	5	private for-profit	0.743
30	Helios Klinik Herzberg/Osterode	9	private for-profit	0.933
31	Kliniken Hochfranken Münchberg	7	public	1.035
32	Kliniken Kreis Mühldorf a. Inn - Klinik Mühldorf	6	public	0.939
33	Helios Klinik Erlenbach	7	private for-profit	0.843
34	Klinikum in den Pfeifferschen Stiftungen gGmbH	6	private non-profit	1.21
35	Klinikum Mittelbaden Rastatt-Forbach	7	public	0.913
36	Klinikum Oberlausitzer Bergland gemeinnützige GmbH	6	public	0.81
37	Helios Klinik Cuxhaven	12	private for-profit	0.964
38	Helios Klinik Jerichower Land	9	private for-profit	0.879
39	Krankenhaus St. Josef Schweinfurt	7	private non-profit	0.854
40	Krankenhaus St. Joseph-Stift Dresden	9	private non-profit	0.94
41	Krankenhaus St. Marienwörth	9	private non-profit	0.858
42	Krankenhaus-Spital Waldshut-Tiengen	9	public	0.761
43	Rhön-Kreisklinik gGmbH Bad Neustadt a.d.Saale	6	private for-profit	0.872
44	Kreiskrankenhaus Emmendingen	9	public	0.886
45	Kreiskrankenhaus Winsen	5	public	0.914
46	Lahn-Dill-Kliniken Dillenburg-Herborn	8	public	1.068
47	Malteser Krankenhaus St. Johannes-Stift Duisburg	10	private non-profit	0.985
48	Maria-Hilf-Krankenhaus Bergheim	11	private non-profit	0.841
49	Marienkrankenhaus Soest	10	private non-profit	1.141
50	Helios Kliniken Mittelweser	10	private for-profit	0.968
51	Paracelsus Klinik Adorf	4	private for-profit	1.091
52	Paracelsus Klinik Schöneck	6	private for-profit	1.091
53	Pleißental-Klinik	5	public	0.851
54	Sankt Marien-Hospital-Buer	9	private non-profit	1.035
55	Segeberger Kliniken GmbH	4	private for-profit	1.457
56	Helios St. Elisabeth-Krankenhaus Bad Kissingen	10	private for-profit	0.864
57	St. Elisabeth-Stift Damme	8	private non-profit	0.897
58	St. Josef -Krankenhaus Engelskirchen kath. Kliniken Oberberg	5	private non-profit	1.133
59	St. Josef-Hospital GFO Kliniken Bonn	7	private non-profit	0.906
60	St. Josef-Krankenhaus Haan	5	private non-profit	1.001
61	St. Josefs-Hospital Cloppenburg gGmbH	9	private non-profit	1.009
62	St. Josefskrankenhaus Heidelberg	6	private non-profit	1.126
63	St. Marien-Krankenhaus Lankwitz	7	private non-profit	1.159
64	St. Nikolaus Stiftshospital GmbH Andernach	10	private non-profit	0.902
65	St. Theresien-Krankenhaus Nürnberg	12	private non-profit	0.907
66	St. Walburga-Krankenhaus Meschede	9	private non-profit	0.846
67	Vinzenz-Pallotti-Hospital	6	private non-profit	0.83
68	Waldkrankenhaus St. Marien Erlangen	11	private non-profit	1.231
69	Westpfalz-Klinikum GmbH Standort Kusel	11	public	1.132
70	Wilhelm Anton Hospital Goch	5	private non-profit	0.843

Table 1: DMUs forming the data sample; source: Klauber et al. (2019)[1]

[1] Note that most data sources are in German. An online translator could be used for translation. However, all necessary data are provided and the usage of the original sources is not necessary.

Apart from the number of departments, Table 1 inherits the ownership type and the case mix index (CMI) of the hospitals. The ownership type and funding is a distinctive feature for hospitals in Germany. 28.8% of all hospitals are public hospitals, 37.1% private-for-profit, and 34.1% private-non-profit (Klauber et al. 2019). It can be argued that the inclusion of different ownership types also results in an inhomogeneous data sample. However, you think the service process in all groups is sufficiently comparable, as hospitals of all types use the same resources (e.g., beds, physicians, and nurses) to deliver the same outputs (treated patients). The CMI reflects the complexity and resource need of all treated cases in the hospital. A low CMI indicates a hospital is treating on average cases of lesser severity, while the opposite is valid for a high CMI. A CMI of 1 reflects the average severity of cases in Germany (Geissler et al. 2011). Including this measure into the analysis seems reasonable for you as, e.g., a complicated heart surgery consumes by far more resources than a standard delivery without complications. From chapter nine of Ozcan (2014), you learn that your feeling is correct, and the adjustment of the number of patients by the CMI has emerged as a standard procedure. In doing so, the analysis as well procures for the fact that a hospital receives more reimbursement for complex patients.

To figure out, which data you need from these hospitals, you consider a recent literature review on DEA in healthcare with a focus on hospitals (Kohl et al. 2018). It gives you an insight into standard input/output settings for hospital DEA studies. You learn that the parameters used most often in hospital DEA studies are beds, nurses, physicians, inpatients, and outpatients. These measures seem fitting for you to describe the service process of a hospital. The number of beds provides a central figure of the capacity of a hospital. Physicians and nurses are most vital for the hospital's service process. Furthermore, personnel costs are the main cost driver of hospitals, and it is advised to use more than one labor category (Chilingerian & Sherman 2011). Therefore, you plan to include the full-time equivalents (FTE) of physicians and nurses in the data sample. You find all the necessary data at hospital search engines as the BKK-Klinikfinder (https://www.bkk-klinikfinder.de) and gather them in Table 2.

DMU	Beds	Physicians	Nurses	Inpatients	Outpatients	DMU	Beds	Physicians	Nurses	Inpatients	Outpatients
1	280	85.75	181.52	12'018	10'578	36	225	44.64	161.49	8'954	9'589
2	242	84.78	156.32	12'761	28'278	37	214	78.50	120.50	10'358	23'987
3	256	83.25	198.66	14'268	31'588	38	241	60.20	117.80	11'524	11'974
4	230	54.40	184.60	8'817	18'739	39	272	57.34	230.95	13'743	20'394
5	200	55.49	145.97	6'228	7'989	40	240	88.65	251.91	13'890	19'908
6	220	81.03	210.70	13'152	15'244	41	274	59.45	161.04	13'215	22'243
7	253	56.48	156.06	6'643	12'523	42	251	75.94	162.69	13'595	23'682
8	255	71.09	213.02	12'173	24'381	43	225	48.88	157.95	10'606	N/A
9	253	67.64	170.73	10'950	20'913	44	263	68.26	150.50	11'699	16'550
10	275	96.75	247.52	14'874	49'511	45	255	77.57	181.12	14'857	28'801
11	250	84.42	167.48	13'343	992	46	261	51.73	189.03	13'323	17'999
12	250	56.74	160.99	9'797	18'644	47	267	61.50	196.70	8'549	9'447
13	245	69.84	177.11	10'862	17'625	48	205	59.94	143.80	8'465	12'963
14	256	44.83	131.93	10'456	5'785	49	265	87.26	227.51	12'995	N/A
15	251	64.85	201.89	11'332	25'200	50	249	97.80	211.80	16'078	27'815

16	228	77.13	157.46	12'580	N/A	51	275	21.90	91.00	5'028	5'153
17	291	96.18	227.32	16'562	35'028	52	275	22.50	81.20	4'244	5'830
18	210	93.30	177.10	13'226	18'900	53	240	55.62	192.45	10'400	16'627
19	247	55.70	144.00	9'578	7'400	54	257	77.98	150.16	4'167	469
20	264	72.90	161.40	11'105	13'692	55	230	55.80	139.60	8'118	4'193
21	275	66.10	127.40	11'180	16'965	56	225	55.30	141.20	10'089	9'116
22	283	93.20	199.30	16'154	18'920	57	244	82.50	221.05	13'672	48'801
23	279	72.05	180.56	11'773	26'781	58	245	49.56	125.12	7'263	20'930
24	261	62.87	188.04	12'528	19'231	59	236	59.33	164.56	12'647	25'616
25	265	103.57	197.86	14'023	23'255	60	217	43.01	130.11	7'935	9'306
26	247	50.36	114.51	5'471	10'914	61	257	82.81	297.02	13'361	57'718
27	263	42.54	149.82	9'705	18'474	62	249	53.99	155.05	8'393	13'798
28	215	57.49	224.45	10'351	N/A	63	274	59.40	142.18	7'508	18'432
29	248	6.00	146.02	10'983	3'031	64	257	61.63	197.36	11'630	28'651
30	214	76.40	153.00	10'654	18'979	65	276	63.47	249.50	11'729	N/A
31	235	42.88	174.44	10'322	9'016	66	232	43.00	147.94	9'686	11'944
32	267	76.08	268.86	15'472	16'497	67	223	62.41	171.89	11'866	35'519
33	262	68.90	167.90	13'938	19'149	68	290	108.49	312.30	12'759	15'962
34	270	73.49	235.14	9'954	11'408	69	244	69.86	219.21	8'304	15'708
35	260	92.37	213.41	13'896	14'036	70	223	54.06	199.77	9'872	14'295

Table 2: Inputs and outputs of the DMUs

Apart from the standard inputs and outputs, you learn from Kohl et al. (2018) that the inclusion of quality indicators has gained rising attention over the past decade. This sounds as well reasonable to you, as many studies you have seen neglected a proper representation of the hospital's service process. The ultimate goal of every hospital is to cure patients. The mere number of admissions in a hospital, however, reveals nothing about the quality of the treatment. When conducting a hospital DEA, the following example should be kept in mind: A hospital, in which every patient is dying because not a single physician is employed and patients cannot be treated, will be rated 100% efficient if the number of patient admissions is the only output. It is needless to say that this example is highly exaggerated. However, the intuition behind it stays relevant for common settings. As evaluable recovery indicators are not existing, the inclusion of quality indicators is indispensable. You know that the publication of quality indicators is mandatory in Germany since the introduction of the DRG system in 2004 (Tiemann et al. 2012). Since then, the Federal Joint Committee (G-BA), which is under statutory supervision of the Federal Ministry of Health is developing the federal quality assurance program. Three of these quality indicators (QI 2009, QI 50722, and QI 50778) are available in the data sample in Table 3. All three indicators deal with community-acquired pneumonia, which is the infection with the highest mortality in Germany. An insufficient treatment is increasing the mortality rate of affected patients (Institute for quality assurance and transparency in health care 2019). QI 2009 documents if an antimicrobial therapy has been started within the first eight hours after admission. Studies have shown increased survivability of patients with an immediately starting antimicrobial therapy (Houck et al. 2004, Mandell et al. 2007). QI 50722 tracks if the respiration rate is measured at the admission. The patient’s respiration rate is important to assess the severity of the infection and to determine

a treatment plan. Finally, QI 50778 reports the ratio of observed to expected deaths. Studies indicate that the implementation of quality management concerning the disease can reduce its mortality rate (Capelastegui et al. 2004). An obligation to provide documentation for the illness is existing since 2005. Overall, the infection is a) relevant, b) treatment affects the patient's recovery and c) is well documented. With these characteristics, the quality indicators of the infection serve as a good quality proxy. The "institute for quality management and transparency in the health system" (IQTIG) is reporting the quality indicators for the disease. In addition, the information is available via hospital search engines as the BKK-Klinikfinder (https://www.bkk-klinikfinder.de) or the Weisse Liste (https://www.weisse-liste.de).

Apart from the mandatory quality indicators, the Weisse Liste reports general patient satisfaction statistics. The recommendation rate and the satisfaction rates for medical care, nursing care, and organization & service are included in the data sample as additional quality indicators. All seven quality indicators are obtainable in Table 3.

DMU	QI 2009	QI 50722	QI 50778	Recom-mendation	Satisfaction with medical care	Satisfaction with nursing care	Satisfaction with organi-zation and service	DMU	QI 2009	QI 50722	QI 50778	Recom-mendation	Satisfaction with medical care	Satisfaction with nursing care	Satisfaction with organi-zation and service
1	95.9	72.8	4.1	78%	77%	81%	74%	36	94.4	97.4	14.9	79%	83%	81%	80%
2	99.5	100	9.3	83%	85%	82%	76%	37	91.1	98.3	15.1	70%	76%	78%	70%
3	91.5	9.8	14.9	85%	85%	84%	80%	38	98.3	98.5	13.3	72%	79%	76%	72%
4	96.1	99.6	14.7	87%	86%	86%	85%	39	96.9	99.6	10.6	86%	83%	84%	80%
5	91.8	100	19.9	89%	89%	88%	83%	40	91.2	96.4	15.6	94%	91%	89%	88%
6	93.6	93.9	7.5	90%	88%	89%	86%	41	95.1	96.7	10.6	87%	86%	84%	80%
7	96.6	98.1	15.7	83%	84%	82%	79%	42	92	97.7	3.4	69%	76%	77%	73%
8	96.9	98.4	8.6	80%	81%	82%	79%	43	89.9	89.6	6.1	73%	78%	78%	75%
9	91.5	98.8	13.3	79%	80%	81%	80%	44	90.8	100	4.7	79%	82%	82%	79%
10	94.2	94.7	11.2	80%	82%	82%	76%	45	90.7	97.3	3.4	82%	81%	79%	76%
11	95.1	96.8	6.7	85%	86%	85%	81%	46	97.7	100	9.1	73%	78%	78%	73%
12	96.9	80.3	8.9	75%	82%	81%	76%	47	97.6	99.6	2.1	79%	81%	79%	74%
13	93.6	98.8	4.8	77%	79%	80%	72%	48	94.4	98.1	4.6	69%	76%	76%	70%
14	96.5	98.9	18.8	75%	78%	78%	73%	49	96.6	100	8.5	84%	82%	81%	78%
15	N/A	N/A	N/A	81%	84%	82%	76%	50	91.6	91.3	10.2	66%	73%	75%	67%
16	100	100	2.6	81%	82%	81%	77%	51	100	98.8	10.1	88%	86%	86%	85%
17	99.2	99.7	9.9	74%	81%	79%	73%	52	N/A	N/A	N/A	88%	86%	86%	85%
18	97.9	92.7	3.5	68%	79%	74%	71%	53	95.1	90.8	15.1	90%	88%	86%	86%
19	95.7	98.1	10	79%	84%	82%	78%	54	97	96.2	13	79%	79%	79%	73%
20	87.1	98.4	11.5	77%	82%	79%	76%	55	97.3	100	4.5	80%	81%	80%	75%
21	95.7	97.5	5.7	72%	78%	77%	71%	56	97.7	74.6	10.2	74%	78%	78%	74%
22	94.5	96.1	11.5	63%	75%	73%	65%	57	88.6	100	5.6	85%	83%	85%	81%
23	100	93.5	11.8	81%	83%	80%	78%	58	96.4	99.5	7.1	81%	82%	82%	78%
24	98.9	98.5	5.6	81%	83%	84%	77%	59	93.3	99.1	3.2	83%	83%	81%	78%
25	96.4	98.7	6	81%	82%	79%	75%	60	90.3	100	3.6	81%	82%	81%	77%
26	99.4	97.5	9.6	74%	81%	77%	71%	61	99.2	99.6	11.2	77%	81%	81%	78%
27	83.8	98.3	11.4	80%	80%	78%	78%	62	100	96.3	9.9	83%	86%	83%	80%
28	90.5	99.3	8.6	63%	71%	70%	62%	63	97	97.8	14.5	76%	80%	77%	76%
29	93.3	88.8	13	90%	89%	85%	87%	64	100	100	4	80%	83%	81%	76%
30	96.2	97.7	8.7	70%	78%	77%	70%	65	89.9	98	11	80%	83%	82%	78%
31	93.3	98	14.6	90%	87%	87%	86%	66	84.8	99	8.9	69%	75%	76%	71%
32	91	96.5	7.4	79%	83%	83%	79%	67	90.8	100	12.1	75%	80%	77%	72%
33	99.4	99.4	16.3	67%	76%	75%	67%	68	89.5	72.5	8.6	84%	85%	83%	79%
34	92.9	97.8	14	86%	85%	85%	81%	69	90.5	98.2	13.5	74%	79%	77%	71%
35	98	85	5.1	73%	78%	76%	73%	70	90.2	98.3	5.3	78%	79%	78%	73%

Table 3: Quality measures of the DMUs

2 Tasks

The forthcoming tasks aim at the conduction of a meaningful DEA in the German hospital sector. By addressing the tasks, you are guided through a range of essential topics, which have to be considered to provide a reasonable DEA. The topics cover the stages of preprocessing, processing, and postprocessing. Figure 2 provides an overview of the treated topics. Furthermore, it shows which tasks are related to which stage in the conduction of DEA.

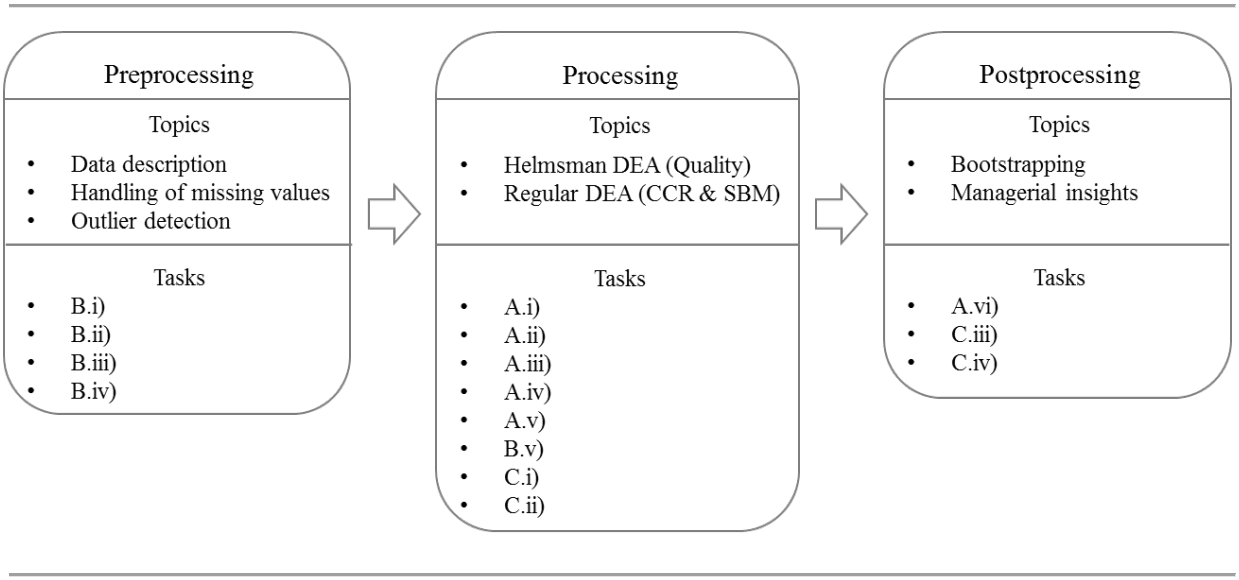


Figure 2: Framework on the conduction of a DEA and the allocation of the tasks to the different stages

A. DEA modeling

i) For the efficiency estimation, a standard CCR model (Charnes et al. 1978) and the more advanced SBM model (Tone 2001) shall be used. Formulate and describe the input-oriented CCR model in its primal (multiplier) and dual (envelopment) form. Explain briefly the linearization of the primal CCR model from its original idea as a fractional program. Provide as well the primal and the dual form of the linearized SBM model. An elaboration on the linearization of the SBM model is not necessary. Please note that standard literature as Cooper et al. (2007) is covering all aspects of this task.

ii) Compare the CCR and SBM model briefly. What is an obvious advantage in the efficiency score calculation of the SBM model over the CCR model?

iii) For the identification of outliers in section B.iv), a super-efficiency DEA model (Andersen & Petersen 1993) is used. Describe the model and its difference to the CCR model briefly.

iv) Think of one reason each, why DEA can be deemed as a fair/unfair estimation method, considering the view of all DMUs. Hint: Consider the multi-input, multi-output setting of the studies.

v) No gold standard for the treatment of quality data in DEA has emerged, yet. The reason for this is manifold. First of all, no standard quality measures for DEA have emerged, yet. Second, consistent quality measures are not always available for a complete data sample. Third, some argue that quality is not a part of technical efficiency (Nuti et al. 2011). If quality measures are used, they are usually treated as additional outputs (Kohl et al. 2018). However, some unresolved issues remain with this procedure. An alternative approach is presented by Ferrier & Trivitt (2013), which is called 'Helmsman'. Describe the approach and its benefits towards the use of quality data as output.

vi) Bootstrapping (Simar & Wilson 1998, 2000b) has emerged as one of the most relevant methodological advancements in the past decade. Describe briefly the idea of bootstrapping in general and why it is useful for DEA. What is the difference between the naive and the smoothed bootstrap method? Hint: The publications of Simar & Wilson are in large parts very technical. Bogetoft & Otto (2011) provide an explanation of the Bootstrap for DEA which is easier to follow.

B. Data description

i) Analyze the data sample and calculate summary statistics (minimum, maximum, mean, standard deviation) for all measures. Provide Boxplot diagrams for the inputs (beds, physicians, nurses) and outputs (inpatients, outpatients). Check if a correlation between the inputs and outputs is present, as demanded by Dyson et al. (2001).

ii) Which dimensions of quality are existing according to Donabedian (1988)? To which dimensions can the provided indicators be assigned?

iii) The data sample reveals missing data for the outputs. That's because the data source reports for some hospital a divergent definition for outpatient cases. In order to avoid distorted results, these values have been left blank. Describe the suggestion of Kuosmanen (2009) on the treatment of missing data and apply it to the data sample.

iv) Data Envelopment Analysis is known to be very sensitive to data errors. To avoid problems, check your data for outliers following the approach of Banker & Chang (2006). Describe briefly the approach and the results. How can the results be explained and what are possible treatments of the issue?

v) When conducting a DEA study, several pitfalls need to be evaded. Among the most common mistakes are the mixture of relative and absolute data and an insufficient number of DMUs compared to the number of inputs and outputs. Describe these issues briefly. What are the protocols Dyson et al. (2001) suggest to tackle them?

C. Results

i) Use the Helmsman DEA approach to create a single quality indicator. Use both, the CCR and SBM model to create the measure and compare the results. Where do the differences in the results come from?

Note: The SBM can't handle values of 0. Replace these values with a marginal value as 1.00E-6.

ii) Multiply the case mix indices (Table 1) with the respective inpatient and outpatient cases of each hospital (Table 2). Afterward, proceed in the same way with the single quality indicators obtained in C.i) to receive quality- and severity-adjusted output figures. Conduct a DEA study with these outputs and the inputs from Table 2 using the CCR and SBM model. Exclude outliers identified in B.iv) from the study. Furthermore, exclude as well DMUs, which receive no meaningful score in any of the models. Compare the results of both models. Which DMUs are deemed efficient and what becomes apparent, when looking at efficient/inefficient units?

iii) What are possible managerial insights for you as a decision maker, if the hospital you are working for is the *Helios Klinik Köthen* (DMU 20)? What are the theoretical suggestions for improvement with regard to the CCR and SBM results? Which hospitals are suitable best practice examples? Note: Standard literature as Cooper et al. (2007) might again be helpful for this task.

What is your feeling on the practicability of these theoretical suggestions? Which implications can be drawn, if you did the same analysis last year and the efficiency score of your hospital has increased by 0.1?

iv) Perform the bootstrap procedure 100 times (using the CCR model), to create confidence intervals and bias-corrected values for your estimates. For the calculation of the bandwidth parameter h, a variety of approaches exist (e.g., Simar & Wilson 2000a, Daraio & Simar 2007, Puenpatom & Rosenman 2008). Use a bandwidth parameter of $h = 0.06158$, which corresponds to the approach of Puenpatom & Rosenman (2008). Compare the results with those of C.ii).

3 Solutions

A. DEA models

i)

The CCR (Charnes, Cooper, Rhodes) model (Charnes et al. 1978) is the initial DEA model. Its general idea is to express (and maximize) efficiency as a ratio of weighted outputs to weighted inputs. The weights are decision variables of the optimization problem and do not need to be predefined. In the fractional programming form, the initial output-to-input ratio idea is palpable:

Parameters & Sets:

n	Number of DMUs
m	Number of Inputs
s	Number of Outputs
$j = 1, \dots, n$	Set of DMUs with index j
$i = 1, \dots, m$	Set of inputs with index i
$r = 1, \dots, s$	Set of outputs with index r
$o \in 1, \dots, n$	DMU under observation
x_{ij}	Input i of DMU j
y_{rj}	Output r of DMU j

Decision variables:

v_i	Weight for input i
u_r	Weight for output r

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

$$\frac{\sum_{r=1}^{s} u_r y_{rj}}{\sum_{i=1}^{m} v_i x_{ij}} \leq 1 \qquad \forall j \tag{1b}$$

$$v_i, u_r \geq 0 \qquad \forall i, r \tag{1c}$$

In the next step, Charnes et al. (1978) linearized their fractional program to facilitate solving the problem.

$$\max \sum_{r=1}^{s} u_r y_{ro} \tag{2a}$$

$$\sum_{r=1}^{s} u_r y_{rj} \leq \sum_{i=1}^{m} v_i x_{ij} \qquad \forall j \tag{2b}$$

$$\sum_{i=1}^{m} v_i x_{io} = 1 \tag{2c}$$

$$v_i, u_r \geq 0 \qquad \forall i, r \tag{2d}$$

This linear program (LP) is known as the multiplier form of the CCR model. As for every LP, a dual formulation for (2) is existing. It is known as the envelopment form and expresses the idea to describe every decision making unit (DMU) as a linear combination of all existing units. If a linear combination exists, that produces the same output with fewer inputs, a DMU is deemed inefficient. Otherwise, it is efficient. This linear combination defines an efficient peer unit for every DMU. The distance of a DMU to its peer unit describes the inefficiency of a unit. As the envelopment form has some computational advantages over the multiplier form and the comparison with the SBM is facilitated, we mainly focus on this form hereafter.

Additional notation:

θ_j	Efficiency score of DMU j
λ_j	Share of DMU j in the efficient peer unit of DMU o

$$\min \theta_o \tag{3a}$$

$$\theta_o x_{io} \geq \sum_{j=1}^{n} \lambda_j x_{ij} \qquad \forall i \tag{3b}$$

$$y_{ro} \leq \sum_{j=1}^{n} \lambda_j y_{rj} \qquad \forall r \tag{3c}$$

$$\lambda_j \geq 0 \qquad \forall j \tag{3d}$$

To receive an efficiency estimate θ_j for every DMU in the data sample, the model needs to be solved n times, with every DMU being once under observation, i.e. $j = o$. Note that $\sum_j \lambda_j$ does not have to be equal to 1 in the CCR model.

The Slacks-Based-Measurement model (SBM), invented by Tone (2001), is built on the idea to designate the input slacks s_i^- from (3b) and the output slacks s_r^+ from (3c) and include them into the efficiency score. Usually, the envelopment form of the model is used, as its idea is only visible in this form. The linear version is given in (4).

Additional notation:

t	Variable used for computational reasons (linearization)
S_{io}^-	Slack in input i of DMU o; $S_{io}^- = t \cdot s_{io}^-$
S_{ro}^+	Slack in output r of DMU o; $S_{io}^+ = t \cdot s_{io}^+$
Λ_j	Share of DMU j in the efficient peer unit of DMU o; $\Lambda_j = t \cdot \lambda_j$

$$\min \theta_o = t - \frac{1}{m}\sum_{i=1}^{m} S_{io}^- / x_{io} \tag{4a}$$

$$t + \frac{1}{s}\sum_{r=1}^{s} S_{ro}^+ / y_{ro} = 1 \tag{4b}$$

$$t x_{io} = \sum_{j=1}^{n} \Lambda_j x_{ij} + S_{io}^- \qquad \forall i \tag{4c}$$

$$t y_{ro} = \sum_{j=1}^{n} \Lambda_j y_{rj} - S_{ro}^+ \qquad \forall r \tag{4d}$$

$$t, \Lambda_j, S_{io}^-, S_{ro}^+ \geq 0 \qquad \forall j, i, r \tag{4e}$$

The dual (multiplier) form of the SBM model is given in (5):

$$\max \theta_o \tag{5a}$$

$$\theta_0 + v_i x_{io} - u_r y_{ro} = 1 \tag{5b}$$

$$\sum_{r=1}^{s} u_r y_{rj} \leq \sum_{i=1}^{m} v_i x_{ij} \tag{5c}$$

$$v_i \geq \frac{1}{m x_{io}} \qquad \forall i \tag{5d}$$

$$u_r \geq \frac{\theta_o}{sy_{ro}} \qquad \forall r \qquad (5e)$$

The model is able to include all slacks into the efficiency score and thus overcomes one deficiency of many other DEA models.

ii)

Both models are linear DEA models that calculate technical efficiency scores. One difference is the orientation of the models. While the CCR model can be input or output oriented, the SBM model has no orientation. An input-oriented model tries to minimize the inputs to reach the given output. An output-oriented model, on the other hand, maximizes the output using the available inputs. Additive models, as the SBM, combine both approaches.

Another difference is the treatment of slacks. While in the CCR model slacks do not contribute to the efficiency score, the SBM score is based on them. This fact is an advantage of the SBM model. In the CCR model, a DMU is deemed efficient, if it reaches a score of 1, and furthermore, no slacks are present (Charnes et al. 1978). This definition, however, leads to misinterpretations of the results, as slacks are rarely reported. In addition, no approach on the interpretation of results with slacks exists. A reliable ranking of DMUs based on CCR scores is therefore not possible.

iii)

The super-efficiency model (Andersen & Petersen 1993) is very close to the CCR model. The only difference is the exclusion of the DMU under consideration in constraints (3b) and (3c) from the reference sets. The result is model (6):

$$\min \theta_o \qquad (6a)$$

$$\theta_o x_{io} \geq \sum_{j=1, j\neq o}^{n} \lambda_j x_{ij} \qquad \forall i \qquad (6b)$$

$$y_{ro} \leq \sum_{j=1, j\neq o}^{n} \lambda_j y_{rj} \qquad \forall r \qquad (6c)$$

$$\lambda_j \geq 0 \qquad \forall j \qquad (6d)$$

From a technical point of view, the modification is equivalent to a restriction of the constraints as solution possibilities are dropped. The results of the underlying minimization problem are therefore at least as high

as those of the CCR model. The practical consequence is the possibility to receive scores larger than 1. This allows discrimination and ranking of efficient units. Units receiving a score larger than 1 are called super-efficient.

iv)

The main problem when considering multi-input, multi-output settings for benchmarking is to create a single measure out of this multi-dimensionality. As a consequence, the inputs and outputs need to be weighted. As every DMU might consider a different set of weights as the ideal solution for themselves, they will not agree on a generally valid set of weights. DEA is a fair estimation method in this regard, as it uses for every DMU its ideal input and output weights. Consequently, no DMU can complain that it would have performed better with other weights. On the other hand, the flexibility to assign the weights freely allows the model to eliminate criteria, where the DMU performs weakly. This can be deemed unfair, as it allows DMUs to avoid the evaluation of crucial criteria.

v)

The Helmsman approach uses DEA to create a single statistic, out of various available measures and was first used by Lovell (1995). The approach abandons the usual DEA setting, where inputs, outputs, and a transformation process in-between are required. Instead, a decision-making apparatus is seen as a single, constant unitary input. Technically, this implies the use of a single input with the value 1 for every DMU. All available measures are handled as outputs (Figure 3).

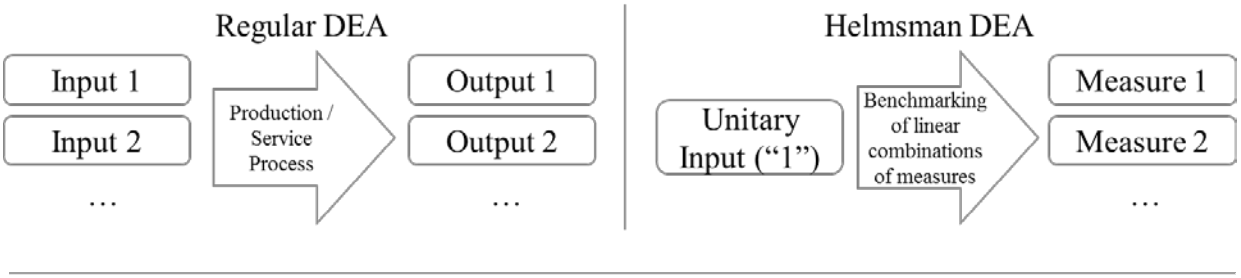


Figure 3: Comparison of a regular DEA and the Helmsman DEA approach

Ferrier & Trivitt (2013) use the Helmsman to create a single quality statistic out of a multitude of quality indicators in a preprocessing stage. Afterward, they multiply the quality measure with the number of inpatients and outpatients, to receive quality adjusted measures. These adjusted outputs are then used in the main DEA study. This procedure is similar to the established approach of multiplying the CMI with the number of treated cases to receive a severity-adjusted measure. The benefit lies in the higher significance of the measure, especially for DEA studies. Having the number of cases and the CMI as separate outputs would allow the DEA model to eliminate one of them by assigning a negligible weight.

As a consequence, highly inefficient hospitals treating inadequately few, but severe cases would be deemed efficient just because of the severity of the cases. A multiplicative adjustment of the cases by the CMI cures the problem and ties the case severity to the number of cases. The same thoughts can be applied to the field of quality. A hospital should neither be deemed efficient by just having high quality without treating an adequate number of patients, nor by treating a lot of patients but with poor quality. Therefore, the Helmsman offers an interesting approach to include multiple factors of quality into DEA analysis. In addition, using the Helmsman to create a single quality indicator in a preceding step prevents the principal DEA analysis from being bloated by too many inputs and outputs. As all indicators are used as outputs, an output-oriented view is reasonable. (7) denotes the resulting model (Ferrier & Trivitt 2013).

$$\max \theta_o \tag{7a}$$

$$\theta_o y_{ro} \leq \sum_{j=1}^{n} \lambda_j \, y_{rj} \qquad \forall r \tag{7b}$$

$$\sum_{j=1}^{n} \lambda_j = 1 \qquad \forall j \tag{7c}$$

$$\lambda_j \geq 0 \qquad \forall j \tag{7d}$$

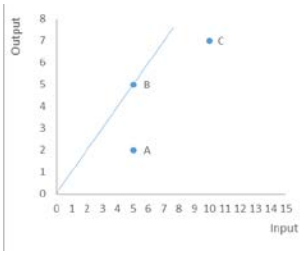
vi)

Bootstrapping is a resampling method promoted for DEA by Simar & Wilson (1998) which constitutes one of the biggest methodological trends in DEA over the past years (Kohl et al. 2018). It is relevant for DEA mainly for two reasons. First, DEA is known to inherit a positive bias in its estimates (Nedelea & Fannin 2013, Mitropoulos et al. 2014). This bias emerges, as the estimated production frontier is based on the DMUs in the data sample. However, not every efficient input/output combination that is theoretically possible is utilized by a DMU in the real world. Therefore, apart from efficient DMUs missing for other reasons, the estimated frontier is always too low (Simar & Wilson 2011). As a consequence, DEA assumes the DMUs to be closer to the production frontier as they really are and assigns efficiency scores which are biased upwards. The bootstrapping procedure can be used to correct for this upwards bias. Second, the possibility to create statistical inference is a huge advancement for DEA studies. The missing knowledge on the robustness of the results is one of the biggest disadvantages of DEA. The bootstrap helps to alleviate the problem with the possibility to create significance intervals for the efficiency estimates.

The general idea of bootstrapping DEA scores is quite simple. It is explained based on an example in Figure 4 with 3 DMUs in a 1 input, 1 output setting. An initial DEA study with the CCR model is conducted

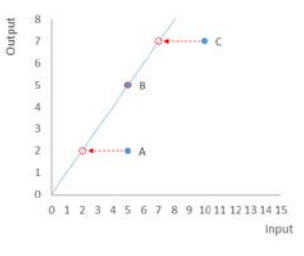
(Step 1 in Figure 4). Afterward, all DMUs are projected to the frontier (Step 2 in Figure 4). In the next step, every DMU j gets a new efficiency score $\theta_j^\star$ assigned. For this new efficiency score, a value β_j from the initial efficiency estimates is randomly drawn with replacement for every DMU j. In the simplest form of bootstrapping, $\theta_j^\star = \beta_j$. This proceeding is called naive bootstrapping. A more sophisticated approach is discussed later on. The projected DMUs are adjusted by their "new" efficiency scores, resulting in a "new" (bootstrapped) data sample (Step 3 in Figure 4). Now, a new DEA study can be conducted in which one by one, every "old" DMU is assessed with regard to the bootstrapped data sample A', B', C'. In step 4 of Figure 4, the result for DMU A is depicted. Afterward, Step 4 is repeated for B and C. The resulting values θ_j^b are the outcome of bootstrap iteration $b = 1$. Steps 3 & 4 are repeated for $b = 1, \dots, B$ bootstrap iterations, with $B = 2000$ being an established scope (Simar & Wilson 2000b).

Step 1



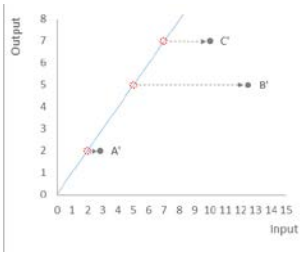
DMU	Input	Output	Efficiency (θ_j)
A	5	2	0.4
B	5	5	1
C	10	7	0.7

Step 2



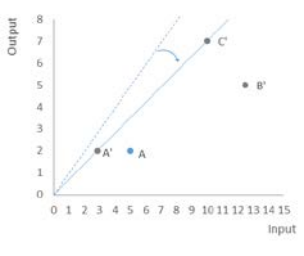
DMU	Projected Input	Projected Output
A	$5 \cdot 0.4 = 2$	2
B	$5 \cdot 1 = 5$	5
C	$10 \cdot 0.7 = 7$	7

Step 3



DMU	New Efficiency ($\theta_j^\star$)	Bootstrapped Input
A'	0.7	$2 \div 0.7 = 2.86$
B'	0.4	$5 \div 0.4 = 12.5$
C'	0.7	$7 \div 0.7 = 10$

Step 4



DMU	Bootstrapped Efficiency (θ_j^b)
A	0.57

Figure 4: Steps in conducting a naive bootstrap iteration for an input-oriented CCR model in an example with 3 DMUs, 1 input, and 1 output

A discussion on the significance and applicability of the naive bootstrap procedure for DEA arose (Löthgren 1998, Ferrier & Hirschberg 1999, Simar & Wilson 1999a, 1999b). The smoothed bootstrap from Simar & Wilson (2000b) emerged as state of the art for DEA applications from this discussion. While the general idea of the procedure is still the same, the resampling of the efficiency scores changes (it is "smoothed"). A problem of the naive procedure is the occurrence of spikes, as only a limited number of efficiency estimates exists from the initial DEA. This problem is growing more severe, the fewer DMUs a study consists of. Kernel density estimators could be used to overcome this issue (see e.g. Silverman 2018). The underlying idea of this method is to distort the naive resampled efficiency scores β_j for DMU j using $\tilde{\theta}_j = \beta_j + h\varepsilon_j$, where ε_j is a standard normal distributed error term. h is a bandwidth parameter, deciding on the magnitude of the smoothing. However, this smoothing procedure needs to be treated with caution, as otherwise, efficiency scores greater than one might be generated. To avoid this pitfall, values bigger than one are reflected (see e.g. Bogetoft & Otto 2011):

$$\tilde{\theta}_j = \begin{cases} \beta_j + h\varepsilon_j & \text{if } \beta_j + h\varepsilon_j \leq 1 \\ 2 - \beta_j - h\varepsilon_j & \text{otherwise} \end{cases} \tag{8}$$

In the next step, the values need to be adjusted to receive parameters with asymptotically correct mean and variance:

$$\theta_j^{\star} = \bar{\beta} + \frac{\tilde{\theta}_j - \bar{\beta}}{\sqrt{1 + h^2/\hat{\sigma}^2}} \tag{9}$$

$\bar{\beta} = \frac{1}{n}\sum_{j=1}^{n}\beta_j$ denotes the mean of the naive bootstrapped efficiency scores and $\hat{\sigma}^2 = \frac{1}{n}\sum_{j=1}^{n}\left(\theta_j - \bar{\theta}_j\right)^2$ the variance of the initial DEA estimates θ_j. The smoothed values $\theta_j^{\star}$ are then used as in the naive bootstrap to create a new (bootstrap) data set, on which the original DMUs are measured against. The result is the efficiency estimate θ_j^b for bootstrap iteration b. In the final step, a bias correction is performed, which reduces the already mentioned overestimation of DEA. An estimator for this bias is calculated via the difference between the average results of the bootstrap iterations and the original DEA estimates:

$$\text{bias}_j^{\star} = \frac{1}{B}\sum_{b=1}^{B}\theta_j^b - \theta_j \tag{10}$$

The bias-corrected bootstrap estimator is calculated by subtracting the bias from the original DEA estimate.

$$\tilde{\theta}_j^{\star} = \theta_j - \left(\frac{1}{B}\sum_{b=1}^{B}\theta_j^b - \theta_j\right) = 2\theta_j - \frac{1}{B}\sum_{b=1}^{B}\theta_j^b \tag{11}$$

The algorithm of Bogetoft & Otto (2011) for the whole procedure can be found in the Appendix.

B. Data description

i)

The dataset contains three inputs (beds, physicians, and nurses) and two outputs (inpatients and outpatients). Furthermore, the CMI and seven quality indicators are present.

Descriptive statistics of all measures can be found in Table 4. An average hospital in the study has 250 beds, 66 physicians and 178.7 nurses. Almost 30,000 patients are treated on average. The average case mix index of 0.97 shows that the hospitals in the study do not only represent an average size for the German hospital market but also treat patients of average severity.

Measure	Min	Max	Mean	Std. Dev.
Beds	200.0	291.0	249.7	21.3
Physicians	6.0	108.5	66.2	19.2
Nurses	81.2	312.3	178.7	44.2
Inpatients	4,167	16,562	11'109.3	2'810.8
Outpatients	469	57,718	18'355.2	10'848.3
CMI	0.74	1.46	0.97	0.1
QI 2009	84	100	94.5	3.8
QI 50722	10	100	94.9	12.1
QI 50778	2	20	9.5	4.3
Recommendation	63%	94%	78.9%	6.9%
Satisfaction with medical care	71%	91%	81.5%	3.9%
Satisfaction with nursing care	70%	89%	80.6%	3.9%
Satisfaction with organization and service	62%	88%	76.5%	5.3%

Table 4: Descriptive statistics of inputs, outputs and quality indicators

A more detailed impression of the inputs and outputs of the data sample can be gained from the boxplot diagrams in Figure 5. Only a few data points are deemed as outliers in the boxplot diagrams, underpinning the homogeneity of the data sample. Even the complete absence of outliers for beds is little surprising as the bed size is fixed to a certain corridor and determines which data is included in the sample. Outstanding is the outlier with regards to physicians. The outlying unit needs less than a tenth of the average number of physicians. In addition, three units are able to treat considerably more outpatients than the rest in the sample. The exclusion of outliers is relevant to prevent the study from distortion. However, outliers cannot be judged by the boxplot diagrams alone. They do not reveal, if in combination with the other measures, a coherent picture emerges, or if unreliable data might be present. An additional test in the upcoming section iv) will be able to answer this question.

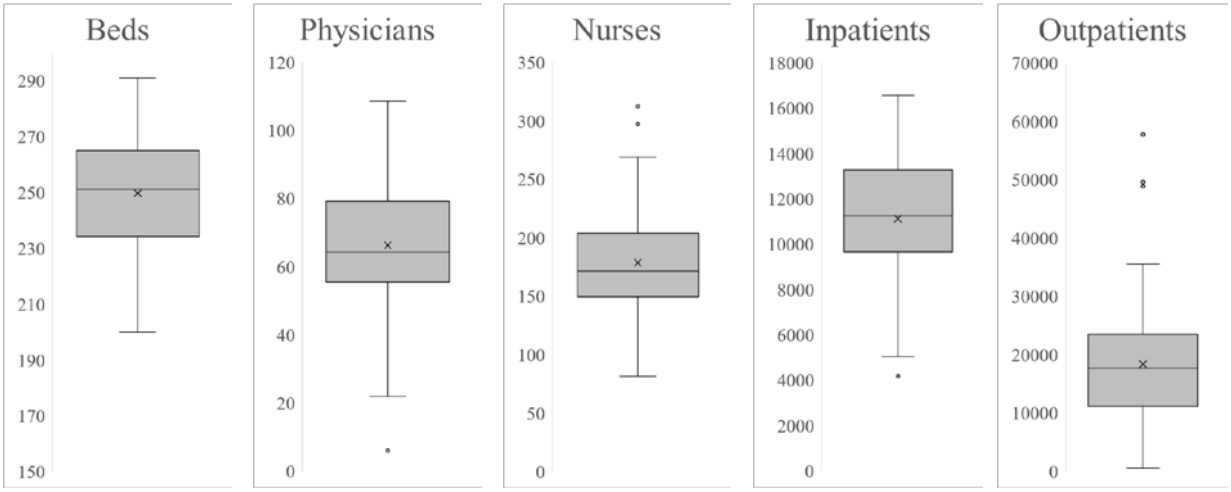


Figure 5: Boxplot diagrams of inputs and outputs

Table 5 displays the correlation between the inputs and outputs. According to Dyson et al. (2001), correlation between inputs and outputs is necessary for DEA to produce meaningful results. As beds are not necessary for the treatment of outpatients, a lower correlation between the two measures is reasonable. However, the number of beds still acts as a proxy for the capacity and endowment of the hospital. Therefore some correlation between the measures is as well not surprising. The observability of the same trend for physicians and nurses is however unexpected. The correlation is in both cases higher for inpatients than for outpatients. Altogether, a significant correlation between inputs and outputs is present.

	Inpatients	Outpatients
Beds	0.259	0.146
Physicians	0.653	0.481
Nurses	0.605	0.519

Table 5: Correlation between the inputs and outputs of the data sample

ii)

With the growing availability of data, the usage of quality measures in DEA became more and more popular. However, Afzali et al. (2009) came to the conclusion, that too little attention has been given to the usage of quality measures in DEA studies, so far. A reason for the unsatisfying use of quality measures in DEA is certainly the difficulty to quantify the abstract term of quality. Furthermore, the term quality includes multiple dimensions in the hospital environment. Donabedian (1988) was the first to analyze the different dimensions of quality of care. He finally defined three dimensions of care:

(1) Structure quality

(2) Process quality

(3) Outcome quality

Structural quality relates to the environment, in which care is executed. It involves material resources, human resources, and organizational structure, as, e.g., methods of reimbursement. Process quality includes actions that are conducted to cure the patient. Here, as well the actions of the patient to seek care, as the diagnosis and treatment by professionals is accounted for. Outcome quality is concerned with the actual result of the treatment. Outcome quality also comprises the satisfaction of a patient with the treatment.

The indicators QI 2009 and QI 50722 are process indicators. QI 50778, the recommendation rate and the satisfaction rates serve as outcome indicators. Note: The structure quality dimension is, according to Ferrier & Trivitt (2013), already covered by the inputs of a DEA.

iii)

Various authors address the field of outlier detection as a preprocessing before the actual DEA analysis with a variety of methods, e.g., Wilson (1993), Simar (2003), Johnson & McGinnis (2008), Bahari & Emrouznejad (2014). Banker & Chang (2006) use the super-efficiency concept to detect outliers. They perform a super-efficiency DEA (Andersen & Petersen 1993) in which the DMU under consideration is excluded from the reference set as described in section A.iii). Therefore, efficiency values bigger than 1 are attainable. According to Banker & Chang (2006), DMUs with an efficiency score larger than 1.2 are possible outliers and should be considered for elimination from the data sample. Table 6 lists the results of the super-efficiency analysis. Nine DMUs receive a score larger than 1. DMU 29, the *Vincentinum Augsburg* is an obvious outlier with a super-efficiency score of 7.11.

DMU	Name	Super-Efficiency Score
29	Klinik Vincentinum Augsburg	7.11
61	St. Josefs-Hospital Cloppenburg gGmbH	1.18
38	Helios Klinik Jerichower Land	1.12
57	St. Elisabeth-Stift Damme	1.10
37	Helios Klinik Cuxhaven	1.08
50	Helios Kliniken Mittelweser	1.06
45	Kreiskrankenhaus Winsen	1.04
67	Vinzenz-Pallotti-Hospital	1.04
41	Krankenhaus St. Marienwörth	1.01

Table 6: DMUs with a super-efficiency score >1

With a look at the data, it is apparent that the unrealistically small number of only six physicians is responsible for the extremely high super-efficiency score. As an explanation serves the exceptionally high number of external physicians in the hospital. As these are not accounted in the number of physicians, an unrealistic super-efficiency score is the results. For this reason, the DMU has to be either excluded from the data sample, the number of external physicians needs to be added to the number of physicians, or an adjustment of the distorted input according to Kuosmanen (2009) need to be carried out. If the number of external physicians is taken into account, the new input value for the physicians of the hospital will result in $6 + 56 = 62$. However, this value bears some distortions as well, as the DMU will be the only one with external physicians being taken into account. Yet, the value might still be relevant as a rather pessimistic estimate. By applying the procedure of Kuosmanen (2009), the value for physicians is replaced by $2 \cdot \max x_{physicians} = 217$. In the results section, the DMU will be excluded from the evaluation.
The other super-efficient units are not that conspicuous, as all have a value smaller than 1.2. Therefore no other unit needs to be taken care of.

iv)
The occurrence of missing or corrupt data in a sample is not rare. Kuosmanen (2009) was the first to address the problem of missing data in DEA systematically. He showed, that the exclusion of DMUs with missing data can distort DEA results more gravely, than the inclusion of the DMU with a replacement value does. Due to DEAs ability to assign the weights freely, a single bad input or output will not influence a DMUs performance seriously. The exclusion of a DMU, however, might change the whole frontier and influence all DMUs that regard the DMU as its peer unit. Therefore, his solution is to assign a pessimistic value to the missing data. For outputs, zero is the most possible pessimistic value. Missing inputs should

take a value M_i that is significantly bigger than the largest value existing in the data sample for input i, i.e. $M_i \gg \max x_i$. Using this procedure, the missing values for outpatients should be replaced with a value of 0.

v)

A mixture of absolute and relative values leads to a distortion of results. While larger units have automatically larger absolute input and output values, indices and relative values are often independent of their size. Dyson et al. (2001) use the following DMUs as an example:

DMU	Input 1	Output 1	Output 2	Output 3
1	10	12	15	1.6
2	20	24	30	1.6

Table 7: Example illustrating the issue of mixing absolute and relative data (Dyson et al. 2001)

Both DMUs work under the same efficiency and environmental conditions and perform equally in the real world. As DMU 2 is twice as big as DMU 1, the values of input 1 and the outputs 1 and 2 are twice the values of DMU 1. Both DMUs have the same value for Output 3, as it is an index measure. The DEA results will be unsatisfying, as only DMU 1 will be deemed efficient. For most DEA models it appears that DMU is producing the same level of Output 3 with half the input.

Note: The CCR model will assign a value of 1 to both DMUs. However, DMU 2 will contain slack on output 3 and can therefore not be deemed efficient. Dyson et al. (2001) advise scaling all measures to work either only with absolute or as relative values.

Another pitfall arises by choosing too many inputs and outputs for a limited amount of DMUs. The more inputs and outputs are chosen, the weaker the discriminative power of DEA. Every additional input or output adds a constraint to the minimization problem (3). Therefore the efficiency scores are inevitably greater or equal with the inclusion of an additional input or output. Dyson et al. (2001) advise to be parsimonious with the number of inputs and outputs and to use at least $2 \cdot (m \cdot s)$ DMUs, where m is the number of inputs and s the number of outputs.

C. Results

i)

In the first stage, DEA is used to create a single quality indicator out of the seven quality indices of the data sample. Both, the CCR and SBM model are used to calculate this Helmsman measure. The results are displayed in Table 8. As the *Vincentinum Augsburg* was identified as an outlier, it will be excluded in the main analysis. Therefore it can be excluded from the Helmsman approach as well. However, as it is no

outlier with regard to the quality indicators, the hospital can as well remain in the sample on this stage to strengthen the data basis.

DMU	Hospital	CCR	SBM	DMU	Hospital	CCR	SBM
1	Agaplesion BathildisKrh. Bad Pyrmont	0.96	0.64	36	Klinikum Oberlausitzer Bergland g GmbH	0.98	0.94
2	Asklepios Klinik Lich GmbH	1.00	1.00	37	Helios Klinik Cuxhaven	0.98	0.86
3	Borromäus-Hospital Leehr gGmbH	0.96	0.42	38	Helios Klinik Jerichower Land	0.99	0.93
4	DIAKOMED Chemnitzer Land	1.00	1.00	39	Krh. St. Josef Schweinfurt	1.00	0.95
5	DiakonieKrh. Halle	1.00	1.00	40	Krh. St. Joseph-Stift Dresden	1.00	1.00
6	DiakonissenKrh. Dresden	1.00	1.00	41	Krh. St. Marienwörth	0.98	0.91
7	Dominikus Krh. GmbH Berlin	1.00	0.99	42	Krh.-Spital Waldshut-Tiengen	0.98	0.56
8	Donau-Ries-Klinik Donauwörth	0.99	0.88	43	RHÖN-Kreisklinik gGmbH Bad Neustadt	0.91	0.70
9	DRK Krh. Luckenwalde	0.99	0.89	44	KreisKrh. Emmendingen	1.00	0.66
10	Elbe Klinikum Buxtehude	0.96	0.88	45	KreisKrh. Winsen	0.97	0.57
11	Ev. DiakonissenKrh. Leipzig	0.98	0.82	46	Lahn-Dill-Kliniken Dillenburg-Herborn	1.00	0.91
12	Ev. Krh. Ludwigsfelde-Teltow	0.97	0.85	47	Malteser Krh. St. Johannes-Stift Duisburg	1.00	0.53
13	Ev. Krh. Mettmann GmbH	0.99	0.67	48	Maria-Hilf-Krh. Bergheim	0.98	0.66
14	Ev. Krh. Bethanien Iserlohn gGmbH	1.00	1.00	49	MarienKrh. Soest	1.00	0.91
15	Gemeinschaftsklinikum Mittelrhein	0.92	0.00	50	Helios Kliniken Mittelweser	0.92	0.78
16	Gesundheitszentrum Tuttlingen	1.00	1.00	51	Paracelsus Klinik Adorf	1.00	1.00
17	Heilig Geist Krh. Köln	1.00	0.95	52	Paracelsus Klinik Schöneck	0.97	0.00
18	Helios Albert-Schweitzer-Kl. Northeim	0.98	0.64	53	Pleißental-Klinik	1.00	1.00
19	Helios Klinik Lutherstadt Eisleben	0.98	0.89	54	Sankt Marien-Hospital-Buer	0.98	0.93
20	Helios Klinik Köthen	0.98	0.85	55	Segeberger Kliniken GmbH	1.00	0.78
21	Helios Klinik Rottweil	0.97	0.73	56	Helios St. Elisabeth-Krh. Bad Kissingen	0.98	0.86
22	Helios St. Marienberg Klinik Helmstedt	0.96	0.82	57	St. Elisabeth-Stift Damme	1.00	0.72
23	Hospital Zum Heiligen Geist Kempen	1.00	1.00	58	St. Josef -Krh. Engelskirchen	0.99	0.84
24	Josephs-Hospital Warendorf	0.99	0.84	59	St. Josef-Hospital GFO Kliniken Bonn	0.99	0.57
25	Katholische Kliniken Ruhrhalbinsel	0.99	0.79	60	St. Josef-Krh. Haan	1.00	0.59
26	St. Marien-Hospital Oberhausen	0.99	0.91	61	St. Josefs-Hospital Cloppenburg gGmbH	1.00	1.00
27	Katholisches Krh. Dortmund-West	0.98	0.85	62	St. JosefsKrh. Heidelberg	1.00	0.97
28	AMEOS Klinik Bremerhaven	0.99	0.73	63	St. Marien-Krh. Lankwitz	0.99	0.95
29	Klinik Vincentinum Augsburg	1.00	0.96	64	St. Nikolaus Stiftshospital Andernach	1.00	1.00
30	Helios Klinik Herzberg/Osterode	0.98	0.83	65	St. Theresien-Krh. Nürnberg	0.98	0.86
31	Kliniken Hochfranken Münchberg	1.00	0.98	66	St. Walburga-Krh. Meschede	0.99	0.77
32	Kliniken Kreis Mühldorf a. Inn	0.97	0.77	67	Vinzenz-Pallotti-Hospital	1.00	0.85
33	Helios Klinik Erlenbach	1.00	1.00	68	WaldKrh. St. Marien Erlangen	0.95	0.78
34	Kl. in den Pfeifferschen Stiftungen gGmbH	0.98	0.94	69	Westpfalz-Klinikum Kusel	0.98	0.86
35	Klinikum Mittelbaden Rastatt-Forbach	0.98	0.74	70	Wilhelm Anton Hospital Goch	0.98	0.68
					Average	0.98	0.82
					Std Dev.	0.02	0.20

Table 8: Single quality indicators based on a CCR and SBM Helmsman approach

The CCR results contain little information. With a minimum of 0.9057, the first quartile at 0.9787, and a standard deviation of 0.02, the discriminative power of the CCR model is extremely limited in this setting. The situation changes with the SBM model. The standard deviation of the results is ten times higher than those of the CCR model. However, a different issue becomes apparent as the SBM model is not able to handle output values of 0. The approach of Kuosmanen (2009) to include DMUs with missing output data is therefore not applicable. When replacing the missing values by a marginal value, these DMUs receive an SBM score of 0. Even a replacement with the minimum value of the respective input among all DMUs is resulting in very low SBM scores of 0.28 for DMU 15 and 0.29 for DMU 52. To prevent these DMUs from misjudgment, they should be excluded from further evaluation.

The differences in the mechanics of the two models become apparent, when looking at the weights, they are assigning. Therefore, the weights of the first ten DMUs are displayed in Table 9.

	CCR							SBM						
DMU	u_1	u_2	u_3	u_4	u_5	u_6	u_7	u_1	u_2	u_3	u_4	u_5	u_6	u_7
1	0.010	0	0	0	0	0	0	0.029	0.001	0.022	0.118	0.119	0.678	0.124
2	0.002	0.004	0	0	0.414	0	0	0.028	0.051	0.015	0.172	0.196	0.174	0.188
3	0.005	0	0.002	0	0.593	0	0	0.001	0.006	0.004	0.070	0.070	0.071	0.075
4	0.004	0	0.006	0	0	0	0.559	0.011	0.001	0.010	0.164	0.166	0.166	0.698
5	0	0	0.050	0	0	0	0	0.002	0.001	0.007	0.161	0.161	0.162	0.172
6	0.003	0	0	0.199	0	0.574	0	0.072	0.002	0.019	0.159	0.162	16.975	0.166
7	0.007	0	0.005	0.044	0.201	0	0	0.021	0.001	0.016	0.171	0.169	0.173	0.179
8	0.002	0.007	0	0	0	0	0.087	0.018	0.001	0.015	0.158	0.156	0.154	0.160
9	0	0.010	0	0	0	0	0	0.001	0.001	0.010	0.161	0.159	0.157	0.159
10	0.003	0.006	0.002	0	0	0.106	0	0.014	0.001	0.011	0.158	0.154	0.154	0.166

u_1 = weight for QI 2009 u_2 = weight for QI 50722 u_3 = weight for QI 11880 u_4 = weight for Recommendation

u_5 = weight for Satisfaction with medical care u_6 = weight for Satisfaction with nursing care u_7 = weight for Satisfaction with organization and service

Table 9: Comparison of the CCR and SBM weights for the Helmsman approach

It should be noted that these DMUs are in large parts representing the findings of the whole data sample, although they are just a random selection. Stunning is the usage of zero weights of the CCR model. 44 of 70 weights are zero for the CCR results in Table 9. For every DMU, at least three weights are zero. DMU 1 is receiving a CCR score of 0.96, although six out of seven weights are zero, and therefore the associated quality indicators are excluded from the evaluation. For the whole data set, 70% of the weights are zeros. For 34 of the 70 DMUs, the CCR model is only taking one single quality indicator into account and

weighing the remaining six with zero. Furthermore, the CCR model is assigning, at least for two inputs of every DMU, a weight of zero. It can be concluded that the multidimensionality of quality is therefore not considered in the single quality indicator if the CCR model is applied. The excessive allocation of zero weights is often mentioned as one of the main drawbacks of the CCR model. The SBM model, on the other hand, is reluctant to assigning weights of zero. This fact explains the higher discriminative power of the SBM model. However, considerable differences still occur and allow emphasizing excellent performance in a specific input. Looking at the weights assigned to QI 50722 ($\widehat{=} u_2$) in Table 9, the differences are obvious. DMU 2 got a weight assigned, which is more than 50 times higher than the weight of most other DMUs. This seems reasonable for DMU 2, as it has the best possible value in QI 50722.

ii)

The main study compares the results from the CCR and SBM model. The single quality indicator and the CMI are used as multiplicative factors for both outputs (e.g. $\text{adjusted inpatients} = \text{CMI} \cdot \text{single quality indicator} \cdot \text{inpatient cases}$). As DMU 29 was identified as an outlier, it is excluded from the analysis. Furthermore, the DMUs 15 and 52 are excluded from the analysis, because they receive no relevant single quality indicator score, due to missing data. Therefore, they have no values for the adjusted inpatients and outpatients and cannot receive a meaningful score in any of the models. The results of the two studies are displayed in Table 10, sorted by SBM scores.

DMU	Hospital	CCR	SBM	DMU	Hospital	CCR	SBM
2	Asklepios Klinik Lich GmbH	1.00	1.00	20	Helios Klinik Köthen	0.75	0.46
40	Krh. St. Joseph-Stift Dresden	1.00	1.00	5	DiakonieKrh. Halle	0.74	0.44
46	Lahn-Dill-Kliniken Dillenburg-Herborn	1.00	1.00	59	St. Josef-Hospital GFO Kliniken Bonn	0.58	0.44
61	St. Josefs-Hospital Cloppenburg gGmbH	1.00	1.00	45	KreisKrh. Winsen	0.58	0.44
10	Elbe Klinikum Buxtehude	0.84	0.78	69	Westpfalz-Klinikum Kusel	0.64	0.44
64	St. Nikolaus Stiftshospital Andernach	0.88	0.78	34	Kl. in den Pfeifferschen Stiftungen gGmbH	0.81	0.40
37	Helios Klinik Cuxhaven	0.90	0.78	26	St. Marien-Hospital Oberhausen	0.54	0.39
17	Heilig Geist Krh. Köln	0.90	0.76	18	Helios Albert-Schweitzer-Kl. Northeim	0.73	0.39
58	St. Josef -Krh. Engelskirchen	0.87	0.73	13	Ev. Krh. Mettmann GmbH	0.53	0.36
67	Vinzenz-Pallotti-Hospital	0.81	0.72	68	WaldKrh. St. Marien Erlangen	0.78	0.36
41	Krh. St. Marienwörth	0.88	0.72	44	KreisKrh. Emmendingen	0.59	0.35
33	Helios Klinik Erlenbach	0.93	0.71	66	St. Walburga-Krh. Meschede	0.61	0.35
63	St. Marien-Krh. Lankwitz	0.84	0.70	32	Kliniken Kreis Mühldorf a. Inn	0.81	0.35
31	Kliniken Hochfranken Münchberg	0.98	0.69	56	Helios St. Elisabeth-Krh. Bad Kissingen	0.71	0.33
7	Dominikus Krh. GmbH Berlin	0.84	0.69	36	Klinikum Oberlausitzer Bergland g GmbH	0.61	0.32
57	St. Elisabeth-Stift Damme	0.74	0.67	42	Krh.-Spital Waldshut-Tiengen	0.45	0.32
51	Paracelsus Klinik Adorf	1.00	0.66	14	Ev. Krh. Bethanien Iserlohn gGmbH	0.87	0.32
62	St. JosefsKrh. Heidelberg	0.82	0.65	19	Helios Klinik Lutherstadt Eisleben	0.76	0.32
9	DRK Krh. Luckenwalde	0.76	0.62	3	Borromäus-Hospital Leehr gGmbH	0.39	0.31

22	Helios St. Marienberg Klinik Helmstedt	0.96	0.61	55	Segeberger Kliniken GmbH	0.88	0.30
50	Helios Kliniken Mittelweser	0.90	0.61	70	Wilhelm Anton Hospital Goch	0.50	0.29
39	Krh. St. Josef Schweinfurt	0.82	0.60	35	Klinikum Mittelbaden Rastatt-Forbach	0.67	0.29
38	Helios Klinik Jerichower Land	0.99	0.59	48	Maria-Hilf-Krh. Bergheim	0.44	0.28
4	DIAKOMED Chemnitzer Land	0.73	0.58	60	St. Josef-Krh. Haan	0.50	0.27
23	Hospital Zum Heiligen Geist Kempen	0.71	0.58	1	Agaplesion BathildisKrh. Bad Pyrmont	0.56	0.24
25	Katholische Kliniken Ruhrhalbinsel	0.82	0.57	47	Malteser Krh. St. Johannes-Stift Duisburg	0.33	0.17
27	Katholisches Krh. Dortmund-West	0.76	0.55	11	Ev. DiakonissenKrh. Leipzig	0.88	0.05
30	Helios Klinik Herzberg/Osterode	0.72	0.54	54	Sankt Marien-Hospital-Buer	0.33	0.02
24	Josephs-Hospital Warendorf	0.73	0.52	49	MarienKrh. Soest	0.96	0.00
8	Donau-Ries-Klinik Donauwörth	0.68	0.50	16	Gesundheitszentrum Tuttlingen	0.94	0.00
21	Helios Klinik Rottweil	0.75	0.49	28	AMEOS Klinik Bremerhaven	0.68	0.00
53	Pleißental-Klinik	0.73	0.49	65	St. Theresien-Krh. Nürnberg	0.65	0.00
12	Ev. Krh. Ludwigsfelde-Teltow	0.65	0.48	43	RHÖN-Kreisklinik gGmbH Bad Neustadt	0.58	0.00
6	DiakonissenKrh. Dresden	0.97	0.48				
					Average	0.73	0.46
					Std Dev.	0.21	0.26

Table 10: CCR and SBM results

Five out of the 67 DMUs reach a CCR score of 1. However, it should be noted that DMU 51 (*Paracelsus Klinik Adorf*) reaches a value of 1.00 only due to rounding to two decimals. Independent from the rounding, DMU 51 cannot be deemed efficient as it, other than the four remaining units, contains input and output slacks. By definition, a unit is only CCR efficient if it reaches a score of one, and no slacks are existing (Charnes et al. 1978). As a consequence, the CCR scores of units containing slacks should not be interpreted. This fact raises questions about the validity of the model, considering that only ten DMUs in the whole study do not report slacks. With a look at the SBM scores, it becomes apparent that all CCR efficient DMUs are as well SBM efficient. Furthermore, the CCR score is always at least as big as the SBM score. Both observations are generally valid (Tone 2001).

For receiving an efficiency score of 1, it is not crucial to have the best value in an input or output. This becomes obvious when looking at the top ten DMUs with regard to the SBM rating, provided in Table 11.

DMU	CCR	SBM	Rank Beds	Rank Phy.	Rank Nur.	Rank A.Inp.	Rank A.Outp.	Rank A.Inp./Beds	Rank A.Inp./Phy.	Rank A.Inp./Nur.	Rank A.Outp./Beds	Rank A.Outp./Phy.	Rank A.Outp./Nur.
2	1.00	1.00	23	59	24	7	5	2	22	**1**	4	9	2
40	1.00	1.00	20	62	67	5	14	**1**	25	31	11	30	32
46	1.00	1.00	47	13	45	6	18	8	**1**	8	20	8	21
61	1.00	1.00	43	56	69	4	**1**	3	13	43	**1**	**1**	**1**
10	0.84	0.78	61	67	65	10	2	15	43	35	2	3	4
64	0.88	0.78	43	32	48	20	6	20	8	28	6	2	9
37	0.90	0.78	4	53	5	35	11	21	51	5	8	20	3
17	0.90	0.76	70	66	61	2	4	10	28	12	7	13	10
58	0.87	0.73	26	11	6	53	12	52	32	25	10	4	5
67	0.81	0.72	9	33	36	37	7	29	37	34	5	5	6

Table 11: Absolute and relative ranking regarding inputs and outputs of the top ten DMUs in the SBM rating

The table shows the absolute ranks for all inputs and outputs. These ranks are scaled in the sense of efficiency analysis, meaning that few inputs and many outputs are positive. As an example, DMU 2 has rank 59 concerning the input “physicians”. This implies that 58 DMUs need fewer physicians than DMU 2. Among the top ten DMUs, only once a DMU receives an absolute rank of 1 for an input or output. DMU 61 is treating the most adjusted outpatients. More important than the absolute performance in inputs and outputs is the relative performance in input/output ratios. All four efficient DMUs are characterized by at least one efficient relation between one input and one output (Table 11). DMU 2 has the best ratio of nurses to inpatients, DMU 40 presents the best ratio between beds and inpatients, DMU 46 the best ratio of physicians to inpatients, and DMU 61 has the best ratio between every input and the output outpatients. Here, the effect of the outstanding performance of DMU 61 with regards to outpatients is getting visible. Overall, excellent performance in one relative ratio is sufficient, to be deemed efficient.

With a look at DMUs that perform extremely poor in the SBM evaluation, DMU 54 (*Sankt Marien-Hospital-Buer*) comes into view. It is the worst unit with outpatient data available. In contrast to DMU 11, which is ranked second to last, both DEA models are on the same page about DMU 54. It is the DMU performing worst in both outputs, although needing a considerable amount of inputs. Therefore it is evident to rate this DMU poorly.

Remarkable is the considerable difference in the evaluation of the two models for many DMUs (Table 12). Especially the DMUs having a replacement value because of missing data are significant. These five DMUs (marked with an asterisk in Table 12) are among the top ten with the most significant difference between CCR and SBM score.

DMU	Name	CCR	SBM	CCR – SBM
49*	MarienKrh. Soest	0.96	0.00	0.96
16*	Gesundheitszentrum Tuttlingen	0.94	0.00	0.94
11	Ev. DiakonissenKrh. Leipzig	0.88	0.05	0.83
28*	AMEOS Klinik Bremerhaven	0.68	0.00	0.68
65*	St. Theresien-Krh. Nürnberg	0.65	0.00	0.65
43*	Rhön-Kreisklinik gGmbH Bad Neustadt	0.58	0.00	0.58
55	Segeberger Kliniken GmbH	0.88	0.30	0.58
14	Ev. Krh. Bethanien Iserlohn gGmbH	0.87	0.32	0.55
6	DiakonissenKrh. Dresden	0.97	0.48	0.49
32	Kliniken Kreis Mühldorf a. Inn	0.81	0.35	0.46

Table 12: DMUs with the highest difference between CCR and SBM score

These results reveal that the procedure of Kuosmanen (2009) for the treatment of missing values is on the one hand not suitable for the evaluation with the SBM model. On the other hand, it is dubious if a unit failing entirely in a dimension is still receiving excellent results (see DMU 49 in the CCR evaluation). Eye-catching is the result of DMU 11. While it receives a satisfying score of 0.88 in the CCR evaluation, the SBM score of 0.05 is deficient. The main reason behind this observation is a weight of zero in the CCR model for the 'outpatient' output. Along with the zero weight comes by far the highest slack value among all DMUs without a replacement value. This slack value does not influence the CCR score. As no DMU contains a slack for the output 'inpatient' and the input slacks are significantly smaller in proportion to the input values, the SBM score on the opposite, is remarkably affected.

iii)

The *Helios Klinik Köthen* (DMU 20) receives a CCR score of 0.75 and an SBM score of 0.46. Both scores stand for the 35th rank of the 67 DMUs and mediocre performance. They imply significant room for improvement. Especially the SBM score of 0.46 sounds alarming. The procedure to identify the shortfalls of a DMU and create an optimal production plan is called projection, as the DMU is projected to the frontier. In an input-oriented CCR model, the input and output values of the projection can be calculated by (12):

$$x_{io} \leftarrow \theta_o x_{io} - s_{io}^- \quad \forall i \tag{12a}$$

$$y_{ro} \leftarrow y_{ro} + s_{ro}^+ \quad \forall r \tag{12b}$$

The projection of SBM-inefficient DMUs works according to the rules of (13):

$$x_{io} \leftarrow x_{io} - s_{io}^{-} \quad \forall i \tag{13a}$$

$$y_{ro} \leftarrow y_{ro} + s_{ro}^{+} \quad \forall r \tag{13b}$$

Note that a retransformation of the slacks from the SBM model (4) is necessary for the interpretation of the slacks in terms of (13a) and (13b). $s_{io}^{-} = S_{io}^{-}/t$ and $s_{io}^{+} = S_{io}^{+}/t$ have to be calculated.
All values that are necessary to conduct the projection analysis for DMU 20 are listed in Table 13:

x_{beds}	$x_{phy.}$	$x_{nur.}$	$y_{A.Inp.}$	$y_{A.Outp.}$
264	72.90	161.40	9'472	11'679

	θ	s_{beds}^{-}	$s_{phy.}^{-}$	$s_{nur.}^{-}$	$s_{A.Inp.}^{+}$	$s_{A.Outp.}^{+}$
CCR	0.75	17.55	0.00	0.00	0.00	6'873.14
SBM	0.46	84.80	12.64	0.00	0.00	19'193.59

Table 13: Data and results of DMU 20

To realize the CCR results, DMU 20 should use $0.75 \cdot 264 - 17.55 = 180.45$ beds, $0.75 \cdot 72.90 = 54.68$ physicians, and $0.75 \cdot 161.40 = 121.05$ nurses to treat $9'472$ inpatients and $11'679 + 6'873.14 = 18'552.14$ outpatients. This means a reduction of 83.76 beds, 18.22 physicians, and 40.35 nurses, while the level of adjusted inpatients needs to be held constant and a significant increase in adjusted outpatients is necessary. The reduction in inputs that is necessary to become SBM efficient can be learned directly from the slacks in Table 13. A direct implementation of this guidance into the hospital's processes is obviously not possible. The *Helios Klinik Köthen* will not be able to treat a similar or even higher number of patients after reducing their inputs in such a significant manner. However, it can have a closer look at the parameters that seem especially affected by inefficiency. Beds on the input side and adjusted outpatients on the output side are sticking out in this regard. Beds are the only input parameter, which reports an additional slack in the CCR model. In the SBM analysis, beds are as well the input with the highest need for adjustment. Beds are not only in absolute terms the input with the highest deficiency. The SBM model suggests a reduction in beds by almost one third, while physicians need only a reduction of around 17%. On the output side, the necessary increase in adjusted outpatients is exceptional and hardly viable. With a CMI of 1.005 and a single quality indicator of 0.85, the adjustment factors of the outpatients are not particularly bad. Although there is as well room for improvement in quality, the mere number of outpatient cases is indeed the main problem. Therefore, raising the number of outpatient cases should be high on the agenda. A comparison with the reference units might explain the vast gap in adjusted outpatients and reveal further interesting management insights. The reference units are those DMUs, with a positive λ_j value and suitable best practice examples for DMU 20. A linear combination of the reference units, weighted with the λ_j-values, is as well another way to calculate the projection for the *Helios Klinik Köthen*. Reference units for the *Helios Klinik Köthen* are *Asklepios Klinik Lich* GmbH ($\lambda_2 = 0.5779$) and

the *Lahn-Dill-Kliniken Dillenburg-Herborn* ($\lambda_{46} = 0.1561$). A more detailed look at the structure and processes of both hospitals should be of high value for the decision makers of the *Helios Klinik Köthen*.
No implications should be drawn by comparing the results with those of other studies. An increased efficiency score can indicate a risen performance. On the other hand, a higher score can also be triggered by a decreasing performance of reference units. To compare the performance of a DMU over several years, special procedures like the Malmquist index (Färe et al. 1994) are necessary.

iv)

DEA estimates are known to be biased upwards (Bogetoft & Otto 2011). Besides the possibility to create confidence intervals for DEA estimates, the bootstrapping procedure allows for bias correction. These bootstrapped results show significantly lower estimates compared to the initial CCR scores (Table 14).

DMU	Hospital	CCR	Boot	DMU	Hospital	CCR	Boot
40	Krh. St. Joseph-Stift Dresden	1.00	0.94	20	Helios Klinik Köthen	0.75	0.70
46	Lahn-Dill-Kliniken Dillenburg-Herborn	1.00	0.85	21	Helios Klinik Rottweil	0.75	0.67
2	Asklepios Klinik Lich GmbH	1.00	0.80	5	DiakonieKrh. Halle	0.74	0.70
61	St. Josefs-Hospital Cloppenburg gGmbH	1.00	0.66	57	St. Elisabeth-Stift Damme	0.74	0.61
51	Paracelsus Klinik Adorf	1.00	0.90	4	DIAKOMED Chemnitzer Land	0.73	0.65
38	Helios Klinik Jerichower Land	0.99	0.92	24	Josephs-Hospital Warendorf	0.73	0.67
31	Kliniken Hochfranken Münchberg	0.98	0.87	53	Pleißental-Klinik	0.73	0.67
6	DiakonissenKrh. Dresden	0.97	0.91	18	Helios Albert-Schweitzer-Kl. Northeim	0.73	0.71
49	MarienKrh. Soest	0.96	0.92	30	Helios Klinik Herzberg/Osterode	0.72	0.67
22	Helios St. Marienberg Klinik Helmstedt	0.96	0.90	56	Helios St. Elisabeth-Krh. Bad Kissingen	0.71	0.67
16	Gesundheitszentrum Tuttlingen	0.94	0.87	23	Hospital Zum Heiligen Geist Kempen	0.71	0.65
33	Helios Klinik Erlenbach	0.93	0.86	28	AMEOS Klinik Bremerhaven	0.68	0.65
37	Helios Klinik Cuxhaven	0.90	0.76	8	Donau-Ries-Klinik Donauwörth	0.68	0.63
50	Helios Kliniken Mittelweser	0.90	0.85	35	Klinikum Mittelbaden Rastatt-Forbach	0.67	0.63
17	Heilig Geist Krh. Köln	0.90	0.83	65	St. Theresien-Krh. Nürnberg	0.65	0.61
55	Segeberger Kliniken GmbH	0.88	0.84	12	Ev. Krh. Ludwigsfelde-Teltow	0.65	0.61
11	Ev. DiakonissenKrh. Leipzig	0.88	0.81	69	Westpfalz-Klinikum Kusel	0.64	0.60
64	St. Nikolaus Stiftshospital Andernach	0.88	0.78	36	Klinikum Oberlausitzer Bergland g GmbH	0.61	0.54
41	Krh. St. Marienwörth	0.88	0.81	66	St. Walburga-Krh. Meschede	0.61	0.55
14	Ev. Krh. Bethanien Iserlohn gGmbH	0.87	0.82	44	KreisKrh. Emmendingen	0.59	0.55
58	St. Josef -Krh. Engelskirchen	0.87	0.78	43	RHÖN-Kreisklinik gGmbH Bad Neustadt	0.58	0.53
7	Dominikus Krh. GmbH Berlin	0.84	0.79	45	KreisKrh. Winsen	0.58	0.53
10	Elbe Klinikum Buxtehude	0.84	0.69	59	St. Josef-Hospital GFO Kliniken Bonn	0.58	0.53
63	St. Marien-Krh. Lankwitz	0.84	0.76	1	Agaplesion BathildisKrh. Bad Pyrmont	0.56	0.52
62	St. JosefsKrh. Heidelberg	0.82	0.76	26	St. Marien-Hospital Oberhausen	0.54	0.49
39	Krh. St. Josef Schweinfurt	0.82	0.73	13	Ev. Krh. Mettmann GmbH	0.53	0.50
25	Katholische Kliniken Ruhrhalbinsel	0.82	0.78	60	St. Josef-Krh. Haan	0.50	0.47
67	Vinzenz-Pallotti-Hospital	0.81	0.71	70	Wilhelm Anton Hospital Goch	0.50	0.47

32	Kliniken Kreis Mühldorf a. Inn	0.81	0.77	42	Krh.-Spital Waldshut-Tiengen	0.45	0.40
34	Kl. in den Pfeifferschen Stiftungen gGmbH	0.81	0.77	48	Maria-Hilf-Krh. Bergheim	0.44	0.41
68	WaldKrh. St. Marien Erlangen	0.78	0.73	3	Borromäus-Hospital Leehr gGmbH	0.39	0.36
19	Helios Klinik Lutherstadt Eisleben	0.76	0.72	47	Malteser Krh. St. Johannes-Stift Duisburg	0.33	0.31
9	DRK Krh. Luckenwalde	0.76	0.69	54	Sankt Marien-Hospital-Buer	0.33	0.30
27	Katholisches Krh. Dortmund-West	0.76	0.67				
					Average	0.73	0.46
					Std Dev.	0.21	0.26

Table 14: Bootstrap results

The average drops from 0.73 to 0.68. Interesting is the correlation between the CCR and the bootstrapped CCR scores. Although still high, a Pearson index of 0.86 indicates a significant difference from perfect correlation. This shows that the bootstrapping procedure does more than merely reducing every estimate by a certain amount. Looking at the difference between the CCR and the bootstrapped CCR scores more closely, differences from -0.02 up to -0.34 arise (Figure 6).

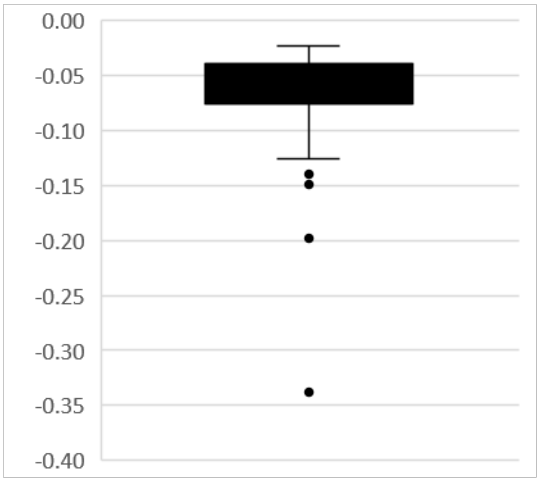


Figure 6: Effect of the bootstrapping procedure on CCR scores (Effect = CCR – Bootstrapped CCR score)

Furthermore, the CCR top performer does not necessarily stay on top after the bootstrapping procedure. While DMU 40 (*Krh. St. Joseph-Stift Dresden*) is still rated best, DMU 61 (*St. Josefs-Hospital Cloppenburg gGmbH*) only receives a bias-corrected score of 0.66 and is the unit with the highest drop. A comparison of the SBM and bootstrapped CCR results shows that the correlation between the SBM and initial CCR scores (0.62) is significantly higher than between the SBM and bootstrapped scores (0.44). The average difference between the SBM and bootstrapped CCR scores (0.23) is slightly smaller than the difference between the SBM and the initial CCR results (0.26). As the bootstrapped CCR scores can be

both, smaller or bigger than the SBM scores, the average absolute differences have been observed as well. The implication, however, does not change. Bootstrapping is a useful addition to the DEA methodology. However, the bootstrap methodology, as a subsequent technique, does not replace the choice of an advanced DEA model.

Notes:

The Bandwidth parameter is calculated by Puenpatom & Rosenman (2008) as $h = 0.9\left(\min\left\{\sigma_\theta, R_{13}/1.34\right\}\right) n^{-1/5}$, where R_{13} is the inter-quartile range of the original DEA estimates and σ_θ denotes the standard deviation of the original DEA estimates.

State of the art is the conduction of 2000 bootstrap iterations. However, to understand the procedure and receive some results, the conduction of 100 bootstrap iterations is sufficient. The conduction of further bootstrap iterations is not providing additional insights and only increases the computational burden.

4 Appendix

4.1 Grading System

The following grading system is suggested for the case study:

Section A	35 Points	Section B	25 Points	Section C	40 Points
i)	8	i)	8	i)	8
ii)	4	ii)	4	ii)	12
iii)	4	iii)	6	iii)	6
iv)	3	iv)	4	iv)	14
v)	6	v)	3		
vi)	10				
					100

Table 15: Grading suggestion for the case study

Note: If the case study is too extensive, the omission of the bootstrapping procedure with the tasks A.vi) and C.iv) is an easy way to reduce its scope. Further modifications are due to the supervisor of the case study.

4.2 Bootstrapping algorithm of Bogetoft & Otto (2011):

(1) Compute θ_j as solution to $\min\{\theta|(\theta x_j, y_j) \in \hat{T}\}$ for $j = 1, \dots, n$

(2) Use bootstrap via smooth sampling from $\theta_1, \dots, \theta_n$ to obtain a bootstrap replica $\theta_1^\star, \dots, \theta_n^\star$. This is done as follows

(2.1) Bootstrap, sample with replacement from $\theta_1, \dots, \theta_n$, and call the results $\beta_1, \dots, \beta_n$

(2.2) Simulate standard normal independent random variables $\varepsilon_1, \dots, \varepsilon_n$

(2.3) Calculate

$$\tilde{\theta}_j = \begin{cases} \beta_j + h\varepsilon_j & \text{if } \beta_j + h\varepsilon_j \leq 1 \\ 2 - \beta_j - h\varepsilon_j & \text{otherwise} \end{cases}$$

Note that by construction, $\tilde{\theta}_j \leq 1$.

(2.4) Adjust $\tilde{\theta}_j$ to obtain parameters with asymptotically correct variance, and then estimate the variance

$\hat{\sigma}^2 = \frac{1}{n}\sum_{j=1}^{n}(\theta_j - \bar{\theta}_j)^2$ and calculate

$$\theta_j^\star = \bar{\beta} + \frac{1}{\sqrt{1 + h^2/\hat{\sigma}^2}}(\tilde{\theta}_j - \bar{\beta})$$

Where $\bar{\beta} = \frac{1}{n}\sum_{j=1}^{n}\beta_j$.

(3) Calculate bootstrapped input based on bootstrap efficiency $x_j^b = \frac{\theta_j}{\theta_j^\star} x_j$.

(4) Solve the DEA program to estimate θ_j^b as

$$\theta_j^b = \min\{\theta \geq 0 | y_j \leq \sum_{j=1}^{n} \lambda_j y_j, \theta x_j \geq \sum_{j=1}^{n} \lambda_j x_j^b, \lambda_j \geq 0, \sum_{j=1}^{n} \lambda_j = 1\} \qquad (j = 1, \dots, n)$$

(5) Repeat the steps from (2.1) to obtain the bootstrap estimates

$$(\theta_1^b, \dots, \theta_n^b) \qquad (b = 1, \dots, B)$$

(6) Calculate the mean and variance of $(\theta_1^b, \dots, \theta_n^b)$ to get the bootstrap estimate $\theta_j^\star$, the bias-corrected estimate $\tilde{\theta}_j^{\star}$, and the variance.

Note that the notation has been adapted slightly to fit the rest of this manuscript. Furthermore, Bogetoft & Otto (2011) use a BCC model (Banker et al. 1984) instead of the CCR model. This results in the additional constraint $\sum_{j=1}^{n} \lambda_j = 1$ in (4).

5 References

Afzali, H. H. A., J. R. Moss, M. A. Mahmood (2009). A conceptual framework for selecting the most appropriate variables for measuring hospital efficiency with a focus on Iranian public hospitals. *Health Services Management Research* **22**(2) 81–91.

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

Bahari, A., A. Emrouznejad (2014). Influential DMUs and outlier detection in data envelopment analysis with an application to health care. *Annals of Operations Research* **223**(1) 95–108.

Banker, R. D., H. Chang (2006). The super-efficiency procedure for outlier identification, not for ranking efficient units. *European Journal of Operational Research* **175**(2) 1311–1320.

Banker, R. D., A. Charnes, W. W. Cooper (1984). Some Models for Estimating Technical and Scale Inefficiencies in Data Envelopment Analysis. *Management Science* **30**(9) 1078-1092.

Bogetoft, P., L. Otto. *Benchmarking with DEA, SFA, and R*. Springer New York, New York, NY.

Capelastegui, A., P. P. España, J. M. Quintana, I. Gorordo, M. Ortega, I. Idoiaga, A. Bilbao (2004). Improvement of process-of-care and outcomes after implementing a guideline for the management of community-acquired pneumonia: a controlled before-and-after design study. *Clinical infectious diseases* **39**(7) 955–963.

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

Chilingerian, J., H.D. Sherman (2011). Health-Care Applications: From Hospitals to Physicians, from Productive Efficiency to Quality Frontiers. In *Handbook on Data Envelopment Analysis*, Cooper, Seiford and Zhu (eds.), Springer US.

Cooper, W. W., L. M. Seiford, K. Tone. *Data envelopment analysis: a comprehensive text with models, applications, references and DEA-solver software*. Springer Science & Business Media.

Daraio, C., L. Simar. *Advanced robust and nonparametric methods in efficiency analysis: Methodology and applications*. Springer Science & Business Media.

Donabedian, A. (1988). The quality of care: how can it be assessed? *Journal of American Medical Association* **260**(12) 1743–1748.

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

Färe, R., S. Grosskopf, B. Lindgren, P. Roos (1994). Productivity Developments in Swedish Hospitals: A Malmquist Output Index Approach. In *Data Envelopment Analysis: Theory, Methodology, and Applications*, Springer Netherlands.

Federal Statistical Office of Germany (2018). Grunddaten der Krankenhäuser, 2017. *Fachserie 12 Reihe 6.1.1*.

Ferrier, G. D., J. G. Hirschberg (1999). Can we bootstrap DEA scores? *Journal of Productivity Analysis* **11**(1) 81–92.

Ferrier, G. D., J. S. Trivitt (2013). Incorporating quality into the measurement of hospital efficiency: a double DEA approach. *Journal of Productivity Analysis* **40**(3) 337–355.

Geissler, A., D. Scheller-Kreinsen, W. Quentin, R. Busse (2011). Germany: Understanding G-DRGs. In *Diagnosis-related groups in Europe. Moving towards transparency, efficiency and quality in hospitals*, Busse, Alexander Geissler, Wilm Quentin and Miriam Wiley (eds.), Maidenhead, England, Open University Press.

Houck, P. M., D. W. Bratzler, W. Nsa, A. Ma, J. G. Bartlett (2004). Antibiotic administration in community-acquired pneumonia. *Chest* **126**(1) 320–322.

Institute for quality assurance and transparency in health care (IQTIG). . Ambulant erworbene Pneumonie. Beschreibung der Qualitätsindikatoren für das Jahr 2017. Retrieved from https://iqtig.org/qs-verfahren/pneu/.

Jacobs, R., P. C. Smith, A. Street. *Measuring efficiency in health care: analytic techniques and health policy*. Cambridge University Press.

Johnson, A. L., L. F. McGinnis (2008). Outlier detection in two-stage semiparametric DEA models. *European Journal of Operational Research* **187**(2) 629–635.

Klauber, J., M. Geraedts, J. Friedrich, J. Wasem (eds.). *Krankenhaus-Report 2019. Das digitale Krankenhaus*, 1st edition. Springer Berlin; Springer, Berlin.

Kohl, S., J. Schoenfelder, A. Fügener, J. O. Brunner (2019). The use of Data Envelopment Analysis (DEA) in healthcare with a focus on hospitals. *Health Care Management Science*, **22**(2), 245-286.

Kuosmanen, T. (2009). Data envelopment analysis with missing data. *Journal of the Operational Research Society* **60**(12) 1767–1774.

Löthgren, M. (1998). How to bootstrap DEA estimators: a Monte Carlo comparison. *Working paper series in Economics and Finance* (223).

Lovell, C. K. (1995). Measuring the macroeconomic performance of the Taiwanese economy. *International Journal of Production Economics* **39**(1-2) 165–178.

Mandell, L. A., R. G. Wunderink, A. Anzueto, J. G. Bartlett, G. D. Campbell, N. C. Dean, S. F. Dowell, T. M. File Jr, D. M. Musher, M. S. Niederman (2007). Infectious Diseases Society of America/American Thoracic Society consensus guidelines on the management of community-acquired pneumonia in adults. *Clinical infectious diseases* **44**(Supplement_2) S27-S72.

Mitropoulos, P., N. Mastrogiannis, I. Mitropoulos (2014). Seeking interactions between patient satisfaction and efficiency in primary healthcare: cluster and DEA analysis. *International Journal of Multicriteria Decision Making* **4**(3) 234–251.

Nedelea, I. C., J. M. Fannin (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.

Nuti, S., C. Daraio, C. Speroni, M. Vainieri (2011). Relationships between technical efficiency and the quality and costs of health care in Italy. *International Journal for Quality in Health Care* **23**(3) 324–330.

Ozcan, Y. A. *Health care benchmarking and performance evaluation: an assessment using Data Envelopment Analysis (DEA)*. Springer Berlin.

Puenpatom, R., R. Rosenman (2008). Efficiency of Thai provincial public hospitals during the introduction of universal health coverage using capitation. *Health Care Management Science* **11**(4) 319–338.

Silverman, B. W. *Density estimation for statistics and data analysis*. Routledge.

Simar, L. (2003). Detecting outliers in frontier models: A simple approach. *Journal of Productivity Analysis* **20**(3) 391–424.

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

Simar, L., P. W. Wilson (1999a). Of Course We Can Bootstrap DEA Scores! But Does It Mean Anything? Logic Trumps Wishful Thinking. *Journal of Productivity Analysis* **11**(1) 93–97.

Simar, L., P. W. Wilson (1999b). Some Problems with the Ferrier/Hirschberg Bootstrap Idea. *Journal of Productivity Analysis* **11**(1) 67–80.

Simar, L., P. W. Wilson (2000a). A general methodology for bootstrapping in non-parametric frontier models. *Journal of Applied Statistics* **27**(6) 779–802.

Simar, L., P. W. Wilson (2000b). Statistical Inference in Nonparametric Frontier Models: The State of the Art. *Journal of Productivity Analysis* **13**(1) 49–78.

Simar, L., P. W. Wilson (2011). Performance of the bootstrap for DEA estimators and iterating the principle. In *Handbook on data envelopment analysis*, Springer.

Tiemann, O., J. Schreyögg, R. Busse (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.

Wilson, P. W. (1993). Detecting outliers in deterministic nonparametric frontier models with multiple outputs. *Journal of Business & Economic Statistics* **11**(3) 319–323.