A Monte Carlo Simulation comparing DEA, SFA and two simple approaches to combine efficiency estimates

Mark Andor* Frederik Hesse**

CAWM Discussion Paper No. 51, University of Münster, September, 2011.

Abstract

In certain circumstances, both researchers and policy makers are faced with the challenge of determining individual efficiency scores for each decision making unit (DMU) under consideration. In this study, we use a Monte Carlo experimentation to analyze the optimal approach to determining individual efficiency scores. Our first research objective is a systematic comparison of the two most popular estimation methods, data envelopment (DEA) and stochastic frontier analysis (SFA). Accordingly we extend the existing comparisons in several ways. We are thus able to identify the factors which influence the performance of the methods and give additional information about the reasons for performance variation. Furthermore, we indicate specific situations in which an estimation technique proves superior. As none of the methods is in all respects superior, in real word applications, such as energy incentive regulation systems, it is regarded as "best-practice" to combine the estimates obtained from DEA and SFA. Hence in a second step, we compare the approaches to transforming the estimates into efficiency scores, with the elementary estimates of the two methods. Our results demonstrate that combination approaches can actually constitute "best-practice" for estimating precise efficiency scores.

Keywords: efficiency, data envelopment analysis, stochastic frontier analysis, simulation, regulation

JEL: C1, C5, D2, L5, Q4

Acknowledgement: We would like to thank Uwe Jensen, participants of the 12th European Workshop on Efficiency and Productivity Analysis in Verona, Italy, and the 11th AIEE European Conference in Vilnius, Lithuania, for their helpful comments and suggestions. Nonetheless, the present work remains our own responsibility, as are any errors and omissions.

* Department of Economic Theory, Westfälische Wilhelms-Universität Münster, Germany.
** Finance Center Münster, Westfälische Wilhelms-Universität Münster, Germany.

1 Introduction

In his classic paper, Farrell (1957) stated that the problem of measuring the efficiency of productivity is important to both economic theorists and economic policy makers. Based on Farrell's work, researchers have developed several methods for measuring efficiency. Despite this progress, after more than five decades of efficiency analysis research, there is still no single superior method.

The efficiency analysis literature can be divided into two main branches: parametric and non-parametric methods. The most important representative of the non-parametric methods is, without doubt, data envelopment analysis (DEA). DEA is a linear programming model originally introduced by Charnes et al. (1978) and extended, amongst others, by Banker et al. (1984) to account for variable returns to scale. DEA develops an empirical frontier function the shape of which is determined by the most efficient producers of the observed dataset. Because efficiency is measured as the distance to this frontier, without considering statistical noise, DEA is a deterministic model. The main advantage of the method is the flexibility due to its non-parametric nature, i.e. no assumption about the production function is required. Parametric methods are based on the econometric ordinary least squares method (OLS). The corrected ordinary least squares method (COLS) estimates the efficient frontier, by shifting the OLS regression towards the most efficient producer. It subsequently measures inefficiency as the distance to this frontier. However, COLS has the same disadvantage as the DEA, it is still deterministic. Aigner et al. (1977) and Meeusen and Broeck (1977) developed a stochastic parametric model, namely stochastic frontier analysis (SFA). SFA is a regression-based approach which integrates two unobserved error terms representing inefficiency and statistical noise. Assuming a production function and specific distributions for the error terms allows calibration via an estimation method (e.g. maximum likelihood). The main advantage is the ability to measure efficiency, while simultaneously considering the presence of statistical noise. The flexibility of DEA and the stochastic

nature of SFA explain why these are the two most popular economic approaches for measuring efficiency.

Due to the fact that the methods usually yield different efficiency scores, researchers and especially policy makers face the problem of determining the "true" efficiency of a sector, individual firms or other decision making units (DMU) like schools, hospitals or universities. Using empirical data, it is impossible to evaluate the performance of the methods, because the "true" efficiency is not known. Monte Carlo simulations are used to avoid this problem. This enables researchers to generate their own artificial dataset under specific assumptions. The performance of the various methods can then be evaluated by comparing the known "true" efficiency with the estimated values. On the basis of this procedure, it is not possible to draw definitive conclusions, because the results are only valid under the specific assumptions. However, it is possible to reveal factors influencing the performance of the methods and to shed some light on the advantages and disadvantages. Consequently, Resti (2000) and Mortimer (2002) conclude that the existing simulation studies neither demonstrate that DEA nor a parametric method has an absolute advantage over their competitors, but the simulation studies do succeed in indicating a range of specific situations in which an estimation technique proves superior.

While there is an abundant literature comparing the two most popular methods, the DEA and the SFA, which use empirical data, simulation studies comparing the two methods by means of cross sectional data are relatively scarce (see Mortimer (2002) for an overview of the existing literature). While the simulation study of Gong and Sickles (1992) uses panel data and focuses more on the choice of functional form and estimation method, Banker et al. (1993) is the first study to analyze the performance of DEA and a stochastic frontier model within a wide range of different settings. Accordingly, Banker et al. (1993) use the moment method, instead of maximum likelihood to estimate the efficiencies, because this method is computationally less demanding. Analogously to the concept of Banker et al. (1993), Ruggiero (1999) and Jensen (2005) use a wide range of settings in their simulation studies, in order to

compare the deterministic COLS and the SFA. Motivated by Ruggieros suggestion (1999), that it would be a useful extension to analyze DEA and SFA across situations not considered in Banker et al. (1993), our first research objective is a systematic comparison of the two methods, using cross sectional data. We thus extend the study of Banker et al. (1993) in three directions. First of all, we apply maximum likelihood, instead of the moment method to estimate the SFA. As computational limits changed over time, the maximum likelihood method is currently the preferred SFA estimation method. Secondly, we extend the scope of values for the influence factors (e. g. Number of DMUs) and add potential influence factors not considered in Banker et al. (1993) (e. g. the input distribution). Thirdly, we consider more performance criteria, and are therefore able to gain a clearer impression of the reasons for performance variation. In a nutshell, the first research objective identifies the most important factors influencing the performance of the different methods and improves the accuracy of information about the reasons for variation.

Our second research objective build on these results and consider the fact that in real-world situations, policy makers know neither the true efficiencies nor the true settings, but often have to set a specific individual efficiency objective for each firm, instead of gaining a degree of understanding of efficiency rankings. In such cases, the individual efficiency score estimation needs to be as robust as possible. For example, all incentive regulation systems for energy markets in Europe apply efficiency estimation methods to determine individual efficiency objectives. Due to the fact, that regulators have no information as to which estimates are closer to the true efficiency, it is seen as "best-practice" to apply several efficiency estimation methods and in a second step, to combine the estimates into firm-specific efficiency objectives (see e.g. Haney and Pollitt (2009)). In addition to this observation of real-world application, in the efficiency analysis literature, researchers also assume, that the use of more than one method could help to avoid the occurrence of "methodological bias" (see, for example, Banker et al. (1994)). Given that, to the best of our knowledge, combination approaches have not yet been analyzed in simulation studies, our second

research objective entails combining the estimated scores. Finally, we are able to determine whether a combination approach is superior to the elementary estimates provided by the methods.

The remainder of this paper is organized as follows. Section 2 describes the general simulation design of the Monte Carlo experiment. In Section 3, we analyze which factors influence the performance of DEA and SFA. Subsequently, we discuss the results of the combination approaches. In the final section we summarise the most important results and provide some directions for further research.

2 Simulation Design

Our simulation design is as follows:

- Variation of sample size (DMU):

  The sample size has already been identified as an important factor of influence on the performance of the various methods. The previous literature generally indicates that sample size influences the performance of both methods, but especially SFA should not be applied to small sample sizes. We extend the range of sample sizes beyond those of Banker et al. (1993): n=15, 20, 25, 30, 40, 50, 75, 100, 150, 200 and 300.

- Variation of the percentage of DMUs on the efficient frontier (PDEF):

  PDEF= 5%, 10% and 30%.

- Variation of collinearity between inputs:

  A further factor considered in studies comparing efficiency methods is the collinearity between the inputs. See for example Jensen (2005), who compared COLS and SFA. Therefore, we successively vary the collinearity between the inputs from no to a high correlation: $\rho(x_1,x_2)$=0, 0.1, 0.25, 0.5, 0.75, 0.9.

- Variation of the moments of input distributions:

  Most of the simulation studies use uniform or normal distributions to generate the inputs. In fact, real world input distributions are usually skewed to the right. For instance, Resti (2000) justifies his use of skewed input distributions, by the fact that there are normally more small and medium-sized companies than large ones and that an unrealistic assumption could influence the performance of the methods. However, in contrast to Resti (2000), we vary the input distribution and are therefore able to evaluate the influence. To the best

of our knowledge, this factor has not been analyzed before.

- Variation of the error term

  The error term is the combination of the inefficiency (σ_u) and the noise (σ_v) terms. The influence of each on the error term is its own standard deviation divided by the overall standard deviation $\lambda = \frac{\sigma_v}{\sigma_u+\sigma_v}$. As it is an inherently important factor for the performance of both methods, in a first step, we analyze the noise and the inefficiency terms separately, changing the ratio λ accordingly. Furthermore, we analyze a simultaneous variation of the absolute values, so that the ratio of both components remains constant.

- Variation of the inefficiency term distribution

  In order to generate the inefficiency term, we use a half normal and a beta distribution. Accordingly, we are able to measure the influence of increasing skewness of the inefficiency, as well as a model misspecification of the SFA.

- Variation of the functional form of the production function

  Intuitively, the production function is the most important part of the data generating process, as it is the instrument used to aggregate the components. Given that its importance is mentioned in many studies, it is notable that most of them focus on only two or three different production functions. By contrast, we use a wide range of production functions, which vary with respect to returns-to-scale and flexibility. Table 1 gives an overview of the twelve production functions, their characteristics and the studies in which they were used.

Nr	PF	Description	Parametrization	Source
I	$\ln(y)=\ln(\beta_0)+\beta_1 \cdot \ln(x_1) + \beta_2 \cdot \ln(x_2)$	Cobb-Douglas, CRS, I.w.d.	β_0=2, β_1=0.4, β_2=0.6	*a*
II		Cobb-Douglas, CRS, I.w.e.	β_0=1, β_1=0.5, β_2=0.5	*b*
III		Cobb-Douglas, CRS, I.w.d.	β_0=1, β_1=0.75, β_2=0.25	
IV		Cobb-Douglas, IRS	β_0=1, β_1=0.6, β_2=0.6	
V		Cobb-Douglas, DRS	β_0=1, β_1=0.4, β_2=0.4	
VI		Cobb-Douglas, Piecewise	for $5 \leq x_1 \leq 10$ and $5 \leq x_2 \leq 10$ β_0=0.631, β_1=0.65, β_2=0.55 for $5 \leq x_1 \leq 10$ and $10 \leq x_2 \leq 15$ β_0=0.794, β_1=0.65, β_2=0.45 for $10 \leq x_1 \leq 15$ and $5 \leq x_2 \leq 10$ β_0=1.259, β_1=0.35, β_2=0.55 for $10 \leq x_1 \leq 15$ and $10 \leq x_2 \leq 15$ β_0=1.585, β_1=0.35, β_2=0.45	*c*
VII	$\beta_0 + \beta_1 \cdot \ln(x_1) + \beta_2 \cdot \ln(x_2) + 0.5 \cdot \beta_{11} \cdot [\ln(x_1)]^2 + 0.5 \cdot \beta_{22} \cdot [\ln(x_2)]^2 + \beta_{12} \cdot \ln(x_1) \cdot \ln(x_2)$	Translog	β_0=1, β_1= β_2=0.3, $\beta_{11} = \beta_{22} = \beta_{12}$ = 0.1	*d*
VIII	$\beta_0 + \beta_1 \cdot \ln(x_1) + \beta_2 \cdot \ln(x_2) + 0.5 \cdot \beta_{11} \cdot [\ln(x_1)]^2 + 0.5 \cdot \beta_{22} \cdot [\ln(x_2)]^2 + \beta_{12} \cdot \ln(x_1) \cdot \ln(x_2)$	Translog	β_0=0.085, β_1 = 0.5, β_2 = 0.44, β_{11} = 0.14, β_{22} = 0.09, β_{12} = -0.22	*e*
IX	$\ln(ye^{\theta\delta})= \ln [\sum_{i=1}^{n} \alpha_i \cdot x_i^{-\rho_i}]^{-\delta/\rho}$	CRESH	θ=0, δ=1, α_1=α_2=0.5, ρ=ρ_i=2	*f*
X			θ=0, δ=1, α_1=α_2=0.5, ρ=ρ_i=0.1	
XI			θ=0, δ=1, α_1=α_2=0.5, ρ=ρ_i=-0.25	
XII			θ=0, δ=1, α_1=α_2=0.5, ρ=ρ_i=-0.67	

Table 1: Variation of production function. CRS: Constant returns to scale; IRS: Increasing returns to scale; DRS: Decreasing returns to scale; I.w.e.: Inputs weighted equally; I.w.d.: Inputs weighted differently. [a] Ruggiero (2007), [b] Adler and Yazhemsky (2010) in modified form, [c] Banker et al. (1994), [d] Cordero et al. (2009), [e] Banker et al. (1994), [f] Yu (1998) in modified form.

2.1 The standard simulation set

Because the structure of the analysis becomes increasingly complex through the integration of all possible combinations, we create a standard setting to reduce this complexity. This standard setting is used as the point of reference for the following sensitivity analysis. We therefore vary the different factors of influence successively, while keeping the remaining parameters of the standard set unchanged. In order to obtain reliable results, each setting is replicated 100 times.

For the standard setting, we follow Ruggiero (1999), Jensen (2005) and others, by using two inputs, x_1 and x_2, which are generated from a uniform distribution with the interval $[5, 15]$. Further, we assume that there is no collinearity between x_1 and x_2. Following Aigner and Chu (1968) and Ruggiero (1999), we assume that the data generating process for 100 DMUs is defined by the following Cobb-Douglas

production function:

$$\ln(y_i) = \underbrace{\ln(2) + 0.4ln(x_{1,i}) + 0.6\ln(x_{2,i})}_{production function} \underbrace{- \ln(u_i) + \ln(v_i)}_{error term} \qquad \text{i=1,...,n.} \qquad (1)$$

where u_i and v_i represent the inefficiency and the statistical noise terms respectively. The noise term v_i is drawn from a normal distribution $v_i \sim N(0, 0.05)$, while the inefficiency term u_i is half-normally distributed $u_i \sim N^+(0, 0.2)$. We do not set a specific percentage of the considered DMUs on the efficient frontier (PDEF=0). The endogenous variable y_i is calculated according to (1). Finally, DEA and SFA are applied to estimate the efficiency scores respectively using $x_{1,i}$ and $x_{2,i}$ and the generated y_i. Table 2 lists the assumptions for the standard set.

Variations	**Standard Set**
Sample size	100
Inputs x_1, x_2	$x_{1,2} \sim U(5, 15)$
Collinearity	0
Noise term	$v_i \sim N(0, 0.05)$
Inefficiency term	$u_i \sim N^+(0, 0.2)$
Percentage on the efficient frontier	0
Production function	PF I see table (1)

Table 2: Overview of the variations in the simulation design

Regarding the methods, it is necessary to choose between the different models for DEA and SFA. For our analysis, we use an output-orientated two-step DEA model with variable returns to scale. Conforming to the usual assumptions for SFA, we assume a Cobb-Douglas production function, a normal distribution for the noise term and a half-normal distribution for the inefficiency term. In contrast to Banker et al. (1993), we use maximum likelihood instead of the moment method as the estimation method.

2.2 Approaches for combining the DEA and SFA estimates

Our second research objective is to compare the results of SFA and DEA with two approaches to combining the efficiency estimates of DEA and SFA. The two approaches are the following:

- 'Best-of-two Method': In a similar manner to the German incentive regulation system for electricity and gas, we calculate the efficiency score using DEA and SFA. The individual efficiency is thus the maximum of both values (see Andor (2009)).

- 'Mean Method': Once again, we calculate the efficiency score using both methods, but instead of using the better one, we now calculate the mean. This approach is, for example, applied in the Finnish incentive regulation system (see Haney and Pollitt (2009)).

2.3 Performance criteria

The evaluation of the methods requires a performance criterion. Ruggiero (1999) and others focus on ranking accuracy, as they use the average rank correlation between the "true" and estimated efficiency. However, in real world applications, ranking accuracy is an inferior performance criterion, because policy makers often have to set individual efficiency objectives. Hence, the ability to measure individual efficiency is the most important factor. Accordingly, we calculate the mean of the absolute deviation (MAD) of the "true" and the estimated efficiency values, and use it as the deciding performance criterion. Nevertheless, we also show the results of the ranking accuracy and discuss them if they are of interest.

In order to gain additional insight into the influence of a particular factor, we give additional information criteria. The MAD yields information about the absolute

deviation, but lacks information about over- and underestimation. Because such information could be useful, we additionally calculate the mean of the deviation (MD), as difference between the "true" and the estimated value. Accordingly, a negative sign indicates that the method on average overestimates the efficiency. However, the MD could lead to misinterpretations when driven by outliers. When only a small number of firms exhibits a large negative deviation, the MD will be negative, although the remainder of the sample has a (small) positive value or vice versa. One way of solving this problem is to calculate the median. However, instead of using the median as a further criterion, we calculate the percentage of underestimated firms (PU), because this percentage is easier to interpret in these circumstances. For example, a PU value of 0.70 implies, that the used method leads to an excessively low efficiency score for 70% of the considered DMUs. We discuss these criteria below, if they yield additional information about the reasons for performance variation.

2.4 Comparison procedure and statistical testing

We focus on two different investigation aspects, the first being inter-comparison. We thus compare the performance of the two methods within a certain setting and test in order to determine, whether the performance levels are significantly different. The second aspect - the intra-comparison - looks at the influence of specific factor variations on the method performance. Here, we compare the performance between the respective setting and the standard set. By doing so, we are able to determine, whether the factor under consideration exerts a significant influence on the performance of the method.

In order to test the differences statistically, in accordance to Banker et al. (1993), we apply the Wilcoxon matched-pairs signed rank (WMP) test with a 95% confidence level. Additionally, we apply the Kolmogorov-Smirnov equality-of-distributions test (KS) with a 95% confidence level to check for differences in the shape of distribution of the MAD. An asterisk indicates a significant difference and a minus indicates

insignificance. Because we use two tests, the order of the symbols is also important. An asterisk followed by a minus sign denotes that the first WMP test indicates significance, while the second KS indicates insignificance. When both test have the same indication, we only show one symbol. Furthermore, we use subscript (superscript) symbols for the inter (intra) comparison. Because, for almost all of our chosen settings, the results of DEA and SFA are significantly different (inter comparison), we only label settings which are insignificant.

3 Results

3.1 Standard set

We now summarize and discuss the results of the simulation study. As the basis for our analysis, we initially consider the results for the standard set. Table 3 presents the performance and information criteria for this set. Concerning the MAD, both methods yield similar results: $\text{MAD}_{SFA} \approx \text{MAD}_{DEA} \approx 0.04$. Despite the fact that Banker et al. (1993) used the moment method to estimate the SFA, both our MAD_{DEA} and MAD_{SFA} are comparable to their results. The MD is below zero for the SFA, indicating that the method generally overestimates. Additionally, the PU indicates that the SFA underestimates 42.1% of the DMUs. The DEA is quite balanced, with a MD ≈0, as well as a PU ≈50%. The final performance criterion of rank correlation is 0.8 for the DEA and 0.87 for the SFA.

		MAD		**MD**		**PU**		**Rank**	
	Set	SFA	DEA	SFA	DEA	SFA	DEA	SFA	DEA
	1	0.037	0.042	-0.012	0.001	42.1%	50.2%	0.866	0.800

Table 3: Performance criteria for DEA and SFA for the standard set

Figure 1 shows the histogram of the 10,000 (100 DMU · 100 simulations) estimated efficiency scores for the DEA and SFA separately, as well as a combined graphic. It illustrates that the distributions of the estimated efficiency scores are quite similar up to a level of 90%. From that level onwards, the main differences become apparent. Because, for each simulation, DEA calculates the efficient frontier subject to the specific input output relations, it is characteristic that a relatively high percentage of the 10,000 DMUs is determined as fully efficient. For each simulation, an average of about 12 of 100 DMUs are on the efficient frontier. For the SFA estimates, it is symptomatic that the distribution is left skewed. Only 0.1% of the DMUs is fully efficient, while a large proportion (50% of the DMUs) is relatively efficient (between 90 and 100%) and a declining fraction is relatively inefficient.

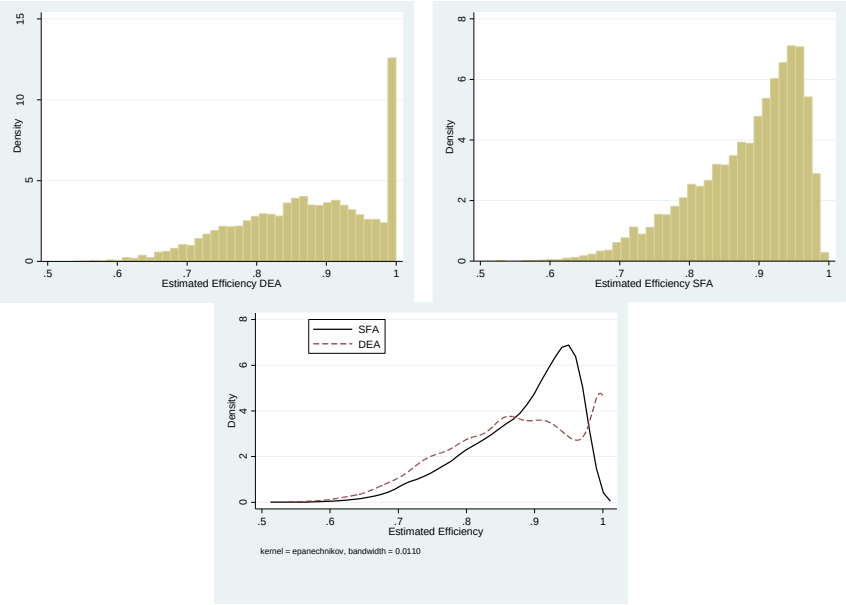

Figure 1: Histogram for SFA and DEA estimates

3.2 Sample size

Table 4 contains the results for the variation of sample size. Regarding the intra comparison, the sample size has a significant influence on both performances. In general, the MAD declines with an increasing number of DMUs and vice versa. Especially for the DEA, it is interesting to note the MD and PU, for which, an increasing sample size leads to both criteria increasing. This can be explained by an decreasing relative number of DMUs on the efficient frontier.

Regarding the inter comparison, one can distinguish between three different intervals. For 50 or more DMUs, the SFA yields a significantly better performance than the DEA. For less than 50, but more than 20 DMUs, the DEA achieves significantly better results. These findings conform to Banker et al. (1993). However, for a further decline in sample size, our results suggest that the SFA performs better than the DEA - the MAD_{SFA} is lower than the MAD_{DEA}. Hence, our results cast doubt on the recommendation of Banker et al. (1993) and Ruggiero (1999), that SFA should not be applied to small sample sizes.

		MAD		MD		PU		Rank	
DMU	Set	SFA	DEA	SFA	DEA	SFA	DEA	SFA	DEA
15	2	0.068*	0.071*	-0.047	-0.063	27.5%	11.7%	0.757	0.641
20	3	0.066*	0.063*	-0.047	-0.052	25.9%	16.9%	0.793	0.670
25	4	0.063*	0.059*	-0.043	-0.045	28.8%	20.7%	0.804	0.666
30	5	0.056*	0.052*	-0.029	-0.033	36.1%	26.6%	0.821	0.706
40	6	0.053*	0.050*	-0.028	-0.027	35.9%	30.0%	0.835	0.737
50	7	0.040*	0.046*	-0.008	-0.019	45.4%	35.4%	0.845	0.751
75	8	0.039^{*-}	0.042^{-}	-0.011	-0.007	43.5%	43.8%	0.863	0.797
100	**1**	**0.037**	**0.042**	**-0.012**	**0.001**	**42.1%**	**50.2%**	**0.866**	**0.800**
150	9	0.034*	0.042^{-}	-0.010	0.013	42.6%	59.6%	0.869	0.822
200	10	0.034*	0.044*	-0.011	0.019	41.7%	64.3%	0.874	0.828
300	11	0.033*	0.046*	-0.012	0.029	39.7 %	71.9%	0.875	0.837

Table 4: Variation of sample size. MAD: mean absolute deviation, MD: mean deviation, PU: percentage of underestimation, Rank: Rankkorrelation. Wilcoxon matched-pairs signed rank test (WMP) and Kolmogorov-Smirnov equality-of-distributions test (KS) with a 95% confidence level. An asterisk followed by a minus sign denotes that the WMP test indicates significance, while the KS indicates insignificance and vice versa. Subscript (superscript) symbols for the inter (intra) comparison. Inter comparison: Only insignificant settings are labeled. See for further details section 2.4.

3.3 Percentage of DMUs on the efficient frontier

The percentage of DMUs on the efficient frontier influences the performance of the methods (see Table 5). With an increasing proportion, the MAD for both methods deteriorate and the probability of underestimation rises. Hence, the SFA is affected more strongly, because it usually predicts only a relatively small number of DMUs as fully efficient. This relative disadvantage of SFA leads to one of the few cases in which the performances no longer differ significantly (PDEF=30%). For higher PDEFs, the DEA is superior to the SFA. The rank correlation also decreases with an increasing PDEF, but in contrast to the MAD, the difference between Rank_{SFA} and Rank_{DEA} is constant at around 0.06. This implies that the SFA still estimates the rank better, but the efficiency scores worse. Hence, the conclusion depends on the considered performance criterion.

		MAD		MD		PU		Rank	
PDEF	Set	SFA	DEA	SFA	DEA	SFA	DEA	SFA	DEA
0%	**1**	**0.037**	**0.042**	**-0.012**	**0.001**	**42.1%**	**50.2%**	**0.866**	**0.800**
5%	12	0.036^{-*}	0.042^{-*}	-0.004	0.010	49.4%	59.1%	0.873	0.811
10%	13	0.039^{*}	0.043^{-*}	0.003	0.017	55.8%	64.9%	0.866	0.810
30%	14	0.046^{*}_{-*}	0.046^{*}_{-*}	0.035	0.032	81.4%	78.6%	0.851	0.803
40%	15	0.056^{*}	0.050^{*}	0.053	0.041	92.0%	85.5%	0.808	0.749
50%	16	0.062^{*}	0.054^{*}	0.059	0.048	94.5%	90.0%	0.766	0.717

Table 5: Variation of the percentage of DMUs on the efficient frontier. MAD: mean absolute deviation, MD: mean deviation, PU: percentage of underestimation, Rank: Rankkorrelation. Wilcoxon matched-pairs signed rank test (WMP) and Kolmogorov-Smirnov equality-of-distributions test (KS) with a 95% confidence level. An asterisk followed by a minus sign denotes that the WMP test indicates significance, while the KS indicates insignificance and vice versa. Subscript (superscript) symbols for the inter (intra) comparison. Inter comparison: Only insignificant settings are labeled. See for further details section 2.4.

3.4 Collinearity

The results for the variation of collinearity between the inputs suggest that it does not exert a significant impact on performance. Table 6 shows that none of the considered criteria are affected. These findings concur with Jensen (2005), who concludes that collinearity has no influence on the performance of SFA and COLS.

		MAD		MD		PU		Rank	
$\rho(x_1,x_2)$	Set	SFA	DEA	SFA	DEA	SFA	DEA	SFA	DEA
0%	**1**	**0.037**	**0.042**	**-0.012**	**0.001**	**42.1%**	**50.2%**	**0.866**	**0.800**
0.1	17	$0.039^{*}-$	0.041^{-}	-0.013	0.005	42.3%	53.0%	0.867	0.813
0.25	18	0.037^{-}	0.042^{-}	-0.010	0.005	43.9%	52.7%	0.871	0.814
0.5	19	0.038^{-}	0.041^{-}	-0.014	0.006	40.9%	53.8%	0.870	0.817
0.75	20	0.036^{*}	0.042^{-}	-0.009	0.011	43.2%	57.6%	0.866	0.820
0.9	21	0.038^{-}	0.043^{-}	-0.014	0.015	40.0%	60.3%	0.862	0.827

Table 6: Variation of collinearity. MAD: mean absolute deviation, MD: mean deviation, PU: percentage of underestimation, Rank: Rankkorrelation. Wilcoxon matched-pairs signed rank test (WMP) and Kolmogorov-Smirnov equality-of-distributions test (KS) with a 95% confidence level. An asterisk followed by a minus sign denotes that the WMP test indicates significance, while the KS indicates insignificance and vice versa. Subscript (superscript) symbols for the inter (intra) comparison. Inter comparison: Only insignificant settings are labeled. See for further details section 2.4.

3.5 Input distribution

As an initial attempt to measure the influence of the input distribution, we vary the distribution without considering specific characteristics. Accordingly, we use the uniform, normal, student-t and gamma distribution successively to generate the inputs. Table 7 gives an overview of the performance of DEA and SFA, when the shape of the input distribution is changing. It is evident that the MAD_{SFA} is not influenced, whereas the MAD_{DEA} changes significantly. Due to these results, we can conclude that the input distribution exerts at least some influence on the performance of the DEA. Below, we determine which characteristics of the distribution exert an influence. Therefore, we vary step-by-step the ratio of standard deviation to mean, the kurtosis and the skewness of the distributions, while holding the other characteristics constant.

		MAD		**MD**		**PU**		**Rank**	
Input(x_1,x_1)	Set	SFA	DEA	SFA	DEA	SFA	DEA	SFA	DEA
***U* [5,15]**	**1**	**0.037**	**0.042**	**-0.012**	**0.001**	**40.5%**	**41.0%**	**0.868**	**0.751**
Normal (150,1)	22	0.038^{-}	0.059^{*}	-0.011	0.055	42.7%	88.2%	0.863	0.873
Student-t(6)	23	0.038^{-}	0.057^{*}	-0.014	0.051	40.3%	85.8%	0.874	0.869
Gamma (1,10)	24	0.039^{-}	0.053^{*}	-0.012	-0.032	42.8%	29.4%	0.866	0.659

Table 7: Variation of input distribution. MAD: mean absolute deviation, MD: mean deviation, PU: percentage of underestimation, Rank: Rankkorrelation. Wilcoxon matched-pairs signed rank test (WMP) and Kolmogorov-Smirnov equality-of-distributions test (KS) with a 95% confidence level. An asterisk followed by a minus sign denotes that the WMP test indicates significance, while the KS indicates insignificance and vice versa. Subscript (superscript) symbols for the inter (intra) comparison. Inter comparison: Only insignificant settings are labeled. See for further details section 2.4.

As already mentioned in the introduction, most simulation studies use uniform or normal distributions to generate the inputs. The normal distribution has an advantage over the uniform distribution regarding parametrization. Because this is useful for our subsequent analysis, we use different normal, instead of uniform distributions.

A change in mean for the normal distribution - keeping the standard deviation constant - changes the *relative* diffusion. For example, the relative diffusion of a

normal distribution with a mean of 10 and a standard deviation of 1 (N (10,1)) is larger than for a normal distribution with the same standard deviation, but a mean of 100 (N(100,1)). We measure this relative diffusion by calculating the ratio of standard deviation and mean (SD/mean).

Concerning the mean and standard deviation, two aspects are of particular interest. Assuming a constant relative diffusion (SD/mean), is the level of mean and standard deviation of relevance? In order to answer this first question, we create group A (setting 25-28) and B (setting 29-32) with a constant ratio of 0.005 and 0.1 respectively, but the settings within the groups differ regarding the level of mean and standard deviation(see Table 8). Accordingly, our results suggest that neither method is influenced, as the MAD within the groups does not change significantly.

But what about the level of the relative diffusion? Instead of comparing the results within the settings, we have to compare the performance between the settings to answer this question. The MAD_{SFA} is not influenced by this kind of variation for both groups, whereas the MAD_{DEA} is - on average - 0.015 worse for group A. A low level of relative diffusion (group A) implies that the inputs are relatively close around the mean. In these cases, the variation of inputs is much smaller, implying a smaller variety of possible outputs of the DMUs, i.e. the range of firm sizes narrows. As a result, the applied DEA with variable returns to scale places fewer DMUs on the efficient frontier. As already mentioned, for the standard setting, around 12 DMUs are, on average, identified as fully efficient by DEA, whereas for group A, only 1.5 DMUs belong to this class. This reduction of fully efficient DMUs may explain the increasing underestimation (PU for group A (B) $\approx$ 88% (63%)). This effect ultimately leads to the deterioration of the MAD_{DEA}. On the other hand, the effect of a declining number of DMUs on the efficient frontier causes an increase in the rank correlation (Rank_{DEA} for group A (B) ≈ 0.87 (≈ 0.83)). Consequently, the performance criteria diverge, yielding different conclusions.

Below, we discuss the two remaining moments - kurtosis and skewness. For the analysis of kurtosis, we use a student-t distribution with different degrees of freedom

			MAD		MD		PU		Rank	
SD/mean	Input(x_1,x_1)	Set	SFA	DEA	SFA	DEA	SFA	DEA	SFA	DEA
0.005	N (800,4)	25	0.038	0.059	-0.015	0.054	40.0%	88.3%	0.860	0.869
0.005	N (400,2)	26	0.038	0.058	-0.012	0.052	41.9%	87.3%	0.866	0.875
0.005	N (200,1)	27	0.039	0.061	-0.017	0.056	39.5%	89.1%	0.866	0.874
0.005	N (100,0.5)	28	0.036	0.058	-0.013	0.053	40.9%	87.8%	0.862	0.870
0.100	N (10,1)	29	0.044	0.044	-0.023	0.019	36.0%	63.0%	0.858	0.823
0.100	N (15,1.5)	30	0.037	0.042	-0.011	0.017	43.1%	62.5%	0.864	0.833
0.100	N (50,5)	31	0.040	0.043	-0.016	0.019	40.1%	64.5%	0.866	0.834
0.100	N (100,10)	32	0.038	0.045	-0.014	0.023	40.5%	66.0%	0.867	0.833

Table 8: Variation of input distribution - SD/mean

(5, 6, 8, and 10). This distribution is advantageous, because the other moments are relatively constant, while the kurtosis is changing. As the benchmark, we use a normal distribution, which is parameterized so that the level of relative diffusion (SD/mean) is similar to the student-t distributions, so as to exclude the influence of a changing SD/mean ratio. Because none of the performance criteria yields substantial changes (see Table 9), we can conclude that the kurtosis does not have an impact on the performance of the methods.

				MAD		MD		PU		Rank	
Kur.	SD/mean	Input(x_1,x_1)	Set	SFA	DEA	SFA	DEA	SFA	DEA	SFA	DEA
3.00	0.010	N (100,1)	33	0.039	0.056	-0.015	0.050	40.7%	86.0%	0.870	0.864
4.00	0.011	*Stud.*-t(10)	34	0.038	0.057	-0.014	0.052	40.1%	86.3%	0.874	0.867
4.50	0.012	*Stud.*-t(8)	35	0.036	0.056	-0.008	0.049	44.5%	85.7%	0.874	0.868
6.37	0.012	*Stud.*-t(6)	36	0.038	0.057	-0.014	0.051	40.3%	85.8%	0.868	0.862
8.02	0.013	*Stud.*-t(5)	37	0.041	0.054	-0.019	0.048	37.7%	84.5%	0.874	0.869

Table 9: Variation of input distribution - kurtosis

For the analysis of skewness, we use two different distributions, one with skewness (gamma), the other one without (uniform), but both with the same level of relative diffusion (see Table 10). As the difference in the kurtosis is negligible, because it does not influence the performance of the methods (see above), a difference in performance should represent the influence of changing skewness. For the two settings with high skewness (settings 41 and 24), it is impossible to create the required comparison settings, because, for a symmetric distribution with, for instance, a level of 0.71 (SD/mean), negative values for the inputs occur.

For all variations of skewness, the SFA is unaffected. For low levels of skewness,

the performance difference of the DEA results within the pairs (1 and 38, 39 and 40) is relatively small, but significantly different. Furthermore, it is evident that for higher levels of skewness, the MAD_{DEA} increases, whereby the deterioration is driven mainly by the increasing overestimation of the DEA (see, for example, set 24: MD_{DEA}=-0.032 and PU_{DEA}=29.4%). The reason could be the application of a DEA VRS. With increasing skewness, the range of firm sizes expands and the number of firms with comparable firm size decreases. Thus, the number of DMUs on the efficient frontier increase (on average 22 (27) for setting 41 (24)) which leads to an increasing overestimation. However, because of the limitation that for high levels of skewness, there is no comparison setting, we can not conclude definitively, that the skewness is the exclusive reason for the increase in MD_{DEA}. However, the results do at least suggest that the skewness has an impact on the performance of DEA.

				MAD		**MD**		**PU**		**Rank**	
Skew.	SD/mean	Input(x_1,x_1)	Set	SFA	DEA	SFA	DEA	SFA	DEA	SFA	DEA
0.00	**0.29**	***U* [5,15]**	**1**	**0.037**	**0.042**	**-0.012**	**0.001**	**42.1%**	**50.2%**	**0.866**	**0.800**
0.63	0.32	G(10,1)	38	0.036^{*-}	0.045^{*}	-0.012	-0.001	41.2%	49.7%	0.866	0.755
0.00	0.43	U[5,35]	39	0.036^{-}	0.043^{-}	-0.012	-0.010	40.8%	42.5%	0.871	0.779
0.89	0.45	G(5,2)	40	0.038^{-}	0.047^{*}	-0.014	-0.009	40.2%	45.8%	0.860	0.723
1.41	0.71	G(2,5)	41	0.037^{-}	0.050^{*}	-0.013	-0.020	41.0%	37.7%	0.861	0.694
2.00	1.00	G(1,10)	24	0.039^{-}	0.053^{*}	-0.012	-0.032	42.8%	29.4%	0.866	0.659

Table 10: Variation of input distribution - skewness. MAD: mean absolute deviation, MD: mean deviation, PU: percentage of underestimation, Rank: Rankkorrelation. Wilcoxon matched-pairs signed rank test (WMP) and Kolmogorov-Smirnov equality-of-distributions test (KS) with a 95% confidence level. An asterisk followed by a minus sign denotes that the WMP test indicates significance, while the KS indicates insignificance and vice versa. Subscript (superscript) symbols for the inter (intra) comparison. Inter comparison: Only insignificant settings are labeled. See for further details section 2.4.

Finally, we can conclude that the input distribution has an influence on the performance of the DEA, but not on the SFA. Our attempts to shed further light on the causes of performance variation, suggest that the kurtosis has no influence, while the skewness and the relationship between standard deviation and mean do exert an influence.

3.6 Standard deviation of the error term

In the following section, we vary the standard deviation of the error term. As the error term consists of both the noise and the inefficiency terms, there are three ways to vary the standard deviation of the noise term: the isolated variation of the standard deviation of the noise and the inefficiency term respectively, which vary according to the ratio $\lambda = \frac{\sigma_v}{\sigma_u+\sigma_v}$, and the simultaneous variation of both standard deviations, whereupon we leave λ constant. Below, we analyze the three options in turn.

The isolated variation of the standard deviation of the noise term (σ_v) yields an unambiguous result. An increasing standard deviation σ_v is combined with an increasing λ and leads to a deteriorating performance of both methods, see Table 11. DEA is affected more by this variation, because it does not account for random noise. While MAD_{SFA} increases from 0.014 to 0.126, MAD_{DEA} rises from 0.028 to 0.195. However, the performance of SFA also diminishes strongly, showing that the separation of noise and inefficiency is less successfull, when the proportion of random noise is high. Accordingly, the percentage of underestimated DMUs increases, so that for σ_v=0.2 and λ=0.5, around PU_{SFA}=72% and PU_{DEA}= 82% of the DMUs are underestimated respectively. The rank correlation also declines for both methods. The results are in line with the findings of Banker et al. (1993).

Table 12 shows the variation of the standard deviation of the inefficiency term. An increasing standard deviation of the inefficiency term decreases λ and the expected efficiency value. In contrast to the variation of the noise term, the methods are affected differently by a variation of the inefficiency. With increasing inefficiency, the DEA improves, while the SFA deterioates. DEA does not account for statistical noise and thus, the results improve with an decreasing ratio of λ, because the proportion of noise reduces. On the other hand, the MAD_{SFA} deteriorates with increasing inefficiency. For cases with a high λ, the skewness of the composed error term is relatively low. As a result of the skewness condition of the SFA, this means that the

			MAD		MD		PU		Rank	
σ_v	λ	Set	SFA	DEA	SFA	DEA	SFA	DEA	SFA	DEA
0.01	4.76%	42	0.014*	0.028*	-0.009	-0.027	31.9%	8.6%	0.985	0.921
0.02	9.09%	43	0.019*	0.028*	-0.007	-0.022	39.9%	23.7%	0.966	0.901
0.05	**20%**	**1**	**0.037**	**0.042**	**-0.012**	**0.001**	**42.1%**	**50.2%**	**0.866**	**0.800**
0.1	33.33%	44	0.060*	0.084*	-0.011	0.054	46.7%	70.0%	0.675	0.630
0.125	38.46%	45	0.065*	0.105*	0.006	0.077	53.2%	74.2%	0.583	0.544
0.15	42.86%	46	0.078*	0.132*	0.016	0.106	57.2%	77.8%	0.516	0.490
0.175	46.66%	47	0.090*	0.155*	0.042	0.132	64.8%	80.4%	0.467	0.440
0.2	50.00%	48	0.105*	0.178*	0.070	0.156	71.5%	82.1%	0.418	0.401

Table 11: Variation of the standard deviation of the noise term (σ_v). MAD: mean absolute deviation, MD: mean deviation, PU: percentage of underestimation, Rank: Rankkorrelation. Wilcoxon matched-pairs signed rank test (WMP) and Kolmogorov-Smirnov equality-of-distributions test (KS) with a 95% confidence level. An asterisk followed by a minus sign denotes that the WMP test indicates significance, while the KS indicates insignificance and vice versa. Subscript (superscript) symbols for the inter (intra) comparison. Inter comparison: Only insignificant settings are labeled. See for further details section 2.4.

SFA recognizes that there is only a small proportion of inefficiency in the data. In these cases, SFA estimates the average inefficiency in the data relatively precisely, but gives every DMU almost the same efficiency score. Hence, the rank correlation is lower, despite a very low MAD. In general, the rank correlation for both methods improves with increasing inefficiency, because, in cases with low inefficiency, the DMUs do not differ much and the estimation of the rank is more difficult. While, for the DEA, both performance criteria lead to the same conclusion, the variation in inefficiency has opposing effects on the MAD and on the rank correlation of the SFA.

After varying the individual standard deviations, in the next step, we change the standard deviation of the noise and inefficiency term simultaneously, so that the relationship between noise and inefficiency remains constant at 20%, as in the standard set. Table 13 shows that the absolute level exerts a diverse influence on the performance criteria. The higher the standard deviation of the error term, the higher the MAD for both methods, while the rank correlation is almost unaffected by this variation. Furthermore, the impact on MD and PU is interesting. On the one hand, the MD_{SFA} and PU_{SFA} decreases with an increasing standard deviation of the error

				MAD		MD		PU		Rank	
σ_u	λ	Exp. Eff.	Set	SFA	DEA	SFA	DEA	SFA	DEA	SFA	DEA
0.02	71.4%	0.984	49	0.023*	0.056*	0.014	0.052	64.8%	79.9%	0.215	0.171
0.05	50.0%	0.962	50	0.027*	0.050*	-0.004	0.038	45.4%	73.0%	0.441	0.381
0.1	33.3%	0.929	51	0.033*	0.044*	-0.008	0.020	46.2%	62.7%	0.692	0.613
0.2	**20.0%**	**0.871**	**1**	**0.037**	**0.042**	**-0.012**	**0.001**	**42.1%**	**50.2%**	**0.866**	**0.800**
0.3	14.3%	0.823	52	0.045*	0.043^{-}	-0.023	-0.010	36.0%	42.8%	0.918	0.873

Table 12: Variation of the standard deviation of the inefficiency term (σ_u). MAD: mean absolute deviation, MD: mean deviation, PU: percentage of underestimation, Rank: Rankkorrelation. Wilcoxon matched-pairs signed rank test (WMP) and Kolmogorov-Smirnov equality-of-distributions test (KS) with a 95% confidence level. An asterisk followed by a minus sign denotes that the WMP test indicates significance, while the KS indicates insignificance and vice versa. Subscript (superscript) symbols for the inter (intra) comparison. Inter comparison: Only insignificant settings are labeled. See for further details section 2.4.

term, while on the other hand, the MD_{DEA} and PU_{DEA} increases. Hence, it can be concluded that the reason for the performance deterioration of both methods are contrary to one another: SFA overestimates and DEA underestimates, with an increasing standard deviation of the composed error term.

				MAD		MD		PU		Rank	
σ_v	σ_u	λ	Set	SFA	DEA	SFA	DEA	SFA	DEA	SFA	DEA
0.025	0.10	20%	53	0.019*	0.024*	-0.004	-0.004	45.5%	43.6%	0.879	0.779
0.0375	0.15	20%	54	0.028*	0.033*	-0.008	-0.002	42.8%	47.1%	0.873	0.793
0.05	**0.20**	**20%**	**1**	**0.037**	**0.042**	**-0.012**	**0.001**	**42.1%**	**50.2%**	**0.866**	**0.800**
0.0625	0.25	20%	55	0.049*	0.050*	-0.024	0.004	36.7%	51.5%	0.854	0.800
0.075	0.30	20%	56	0.056*	0.057*	-0.027	0.010	37.2%	54.5%	0.855	0.808

Table 13: Variation of the absolute level of the error term. MAD: mean absolute deviation, MD: mean deviation, PU: percentage of underestimation, Rank: Rankkorrelation. Wilcoxon matched-pairs signed rank test (WMP) and Kolmogorov-Smirnov equality-of-distributions test (KS) with a 95% confidence level. An asterisk followed by a minus sign denotes that the WMP test indicates significance, while the KS indicates insignificance and vice versa. Subscript (superscript) symbols for the inter (intra) comparison. Inter comparison: Only insignificant settings are labeled. See for further details section 2.4.

3.7 Distribution of the inefficiency term

We now vary the distribution of the inefficiency term in the data generating process, so as to measure its influence on the methods. We thus analyze the influence of the

inefficiency distribution, by comparing the results of a half-normally distributed and a beta distributed inefficiency term. This is particularly interesting, with regard to the SFA. Because we consistently assume a half-normally distributed inefficiency term for the SFA, we are able to analyze the effect of this model specification error. The subsequent comparison is always between a pair of results, containing one setting using the half normal and one using the beta distribution. The parametrization of both settings is chosen in such a manner that they have the same expected value for the efficiency. Regarding the differences in skewness (and kurtosis), we calculate the over-skewness (-kurtosis), representing the skewness (kurtosis) of the beta minus the skewness (kurtosis) of the half normal distribution. Accordingly, a positive over-skewness (-kurtosis) implies that the beta distribution is more skewed (has a higher kurtosis).

The results in Table 14 show a definite tendency. The skewness of the inefficiency distribution has a negative influence on the performance of both methods (regarding the MAD). Furthermore, the results confirm that the SFA is affected primarily, because of the misspecification of the inefficiency distribution. The MD and the PU explain the performance variation: the skewer the distribution, the higher the percentage of underestimated DMUs. In contrast to the MAD, the rank correlation is positively affected (or not affected) by the inefficiency distribution variation.

3.8 Production function

The second possible misspecification error from applying the SFA could arise from assuming an inaccurate production function. The influence of the production function is frequently referred to be important in the literature, but is rarely analyzed. Furthermore, the range of production functions under consideration has been limited so far. For example, Gong and Sickles (1992) use three different production functions, while Banker et al. (1993) use two very similar ones in their simulation studies. We generate the data with a total of twelve different production functions,

					MAD		**MD**		**PU**		**Rank**	
σ_u	$Ex.Ef.$	O.sk.	O.kur.	Set	SFA	DEA	SFA	DEA	SFA	DEA	SFA	DEA
N^+ (0,0.02)	0.984	4.74	40.62	57	0.023^+	0.056^+	0.014	0.052	64.8%	79.9%	0.215	0.171
B (0.065,4)	0.987			58	0.057^+	0.062^+	0.056	0.060	97.0%	89.4%	0.219	0.195
N^+ (0,0.05)	0.962	2.56	14.61	59	0.027^+	0.050^+	-0.004	0.038	45.4%	73.0%	0.441	0.381
B (0.16,4)	0.968			60	0.062^+	0.058^+	0.060	0.054	96.2%	82.4%	0.484	0.439
N^+ (0,0.1)	0.929	1.25	4.84	61	0.033^+	0.044^+	-0.008	0.020	46.2%	62.7%	0.692	0.613
B (0.35,4)	0.935			62	0.057^+	0.052^+	0.054	0.042	91.2%	75.8%	0.721	0.670
N$^+$ (0,0.2)	**0.871**	0.34	0.70	**1**	**0.037**	**0.042**	**-0.012**	**0.001**	**42.1%**	**50.2%**	**0.866**	**0.800**
B (0.75,4)	0.876			63	0.037^+	0.042^-	0.022	0.015	70.6%	59.6%	0.876	0.824
N^+ (0,0.3)	0.822	-0.15	-0.66	64	0.045^+_-	0.043^-_-	-0.023	-0.010	36.0%	42.8%	0.918	0.873
B (1.25,4)	0.822			65	0.045^+_{*-}	0.045^+_{*-}	-0.029	-0.015	29.4%	39.1%	0.916	0.864

Table 14: Variation of the inefficiency term distribution. MAD: mean absolute deviation, MD: mean deviation, PU: percentage of underestimation, Rank: Rankkorrelation. Wilcoxon matched-pairs signed rank test (WMP) and Kolmogorov-Smirnov equality-of-distributions test (KS) with a 95% confidence level. An asterisk followed by a minus sign denotes that the WMP test indicates significance, while the KS indicates insignificance and vice versa. Subscript (superscript) symbols for the inter (intra) comparison. Inter comparison: Only insignificant settings are labeled. See for further details section 2.4.

which vary with respect to returns-to-scale and flexibility.

		MAD		**MD**		**PU**		**Rank**	
PF	Set	SFA	DEA	SFA	DEA	SFA	DEA	SFA	DEA
I	**1**	**0.037**	**0.042**	**-0.012**	**0.001**	**42.1%**	**50.2%**	**0.866**	**0.800**
II	66	0.039^*	0.042^-	-0.015	0.002	40.5%	50.4%	0.863	0.804
III	67	0.039^*	0.042^-	-0.014	0.006	40.5%	53.1%	0.859	0.801
IV	68	0.041^*	0.043^*	-0.015	0.004	41.0%	52.4%	0.863	0.799
V	69	0.039^-	0.041^{*-}	-0.015	0.000	41.6%	48.7%	0.867	0.806
VI	70	0.041^*	0.040^*	-0.014	-0.003	42.5%	46.0%	0.851	0.821
VII	71	0.040^*	0.057^*	-0.015	0.026	41.1%	62.2%	0.861	0.758
VIII	72	0.043^*	0.044^*	-0.017	0.010	40.6%	57.2%	0.840	0.799
IX	73	0.047^*	0.041^*	-0.008	-0.001	46.2%	46.9%	0.788	0.812
X	74	0.036^-	0.042^-	-0.011	0.002	41.3%	50.8%	0.867	0.811
XI	75	0.038^-	0.043^-	-0.012	0.004	42.9%	52.7%	0.863	0.800
XII	76	0.040^*	0.043^-	-0.016	0.006	39.8%	54.8%	0.853	0.800

Table 15: Variation of production function. MAD: mean absolute deviation, MD: mean deviation, PU: percentage of underestimation, Rank: Rankkorrelation. Wilcoxon matched-pairs signed rank test (WMP) and Kolmogorov-Smirnov equality-of-distributions test (KS) with a 95% confidence level. An asterisk followed by a minus sign denotes that the WMP test indicates significance, while the KS indicates insignificance and vice versa. Subscript (superscript) symbols for the inter (intra) comparison. Inter comparison: Only insignificant settings are labeled. See for further details section 2.4.

In Table 15, the results for the variation of the production function are presented. In contrast to the statements made in the literature, the results suggest that the

production function has - in relation to other influence factors - a relatively weak influence on the performance of the methods. For the majority of variations, the production function seems to have no relevant influence, even though there is a significant performance difference. The differences between the MAD of the standard set and the MAD of these settings are between 0.002 for the DEA and 0.006 for the SFA. Yet, in some cases, this can have a crucial effect. The performance of the SFA is relatively worse (MAD_{SFA}=0.047), when the data is generated from a CRESH production function with ρ=2.0 (PF IX). The reason could be that this production function is characterized by a relatively low elasticity of substitution (0.333), in contrast to the assumed Cobb-Douglas function, which has an elasticity of substitution of 1. The DEA performance is also relevantly influenced in one case. If the data is generated from PF VII, which is a specific translog production function (see Table 1), the MAD_{DEA} rises to 0.057. In summary, our results suggest that the influence of the production function should not be overrated, but can in certain cases, effect the performance of the methods.

4 Comparison of approaches for determining individual efficiency scores

Our second research objective considers the fact that, in real world applications, it is regarded as "best-practice" to apply several efficiency estimation methods and to combine the achieved estimates into firm-specific efficiency objectives (see e.g. Haney and Pollitt (2009)). In addition to this observation of real-world application, also in the efficiency analysis literature, researchers assume that the use of more than one method could help to avoid the possible occurrence of "methodological bias" (see for example Banker et al. (1994)). Below, we compare two combination approaches, the 'Best of two' (BOT) and the Mean Method (MM), with the elementary estimates of DEA and SFA. For the analysis, we use the same settings as for the isolated analysis of DEA and SFA, but do not analyze the influence of parameter variations on the performance in the same detail as in the previous section. Rather, we present the results at an aggregated level, so as to concentrate on the comparison of the different approaches to setting individual efficiency objectives.

Table 16 presents the average performance criteria for all 76 settings. The results confirm that a transformation approach can be superior to the elementary estimates. As shown in Table 16, MM has the lowest MAD, even though the SFA is very close to this value (MAD_{SFA}=0.043, MAD_{MM}=0.042). Regarding the rank correlation, the SFA has the highest value, followed by the MM. Hence, a definite conclusion is not possible. Comparing all settings, the DEA is clearly the poorest method. Bearing in mind that the standard setting favors the SFA (e.g. with respect to the inefficiency distribution assumption), it can be assumed that under different assumptions, the relative performance of the DEA improves. In such cases, we expect a further relative performance enhancement of the MM. Further research could investigate this expectation.

Focusing on the combination approaches, it is evident that the MM is superior to

BOT, as both performance criteria (MAD_{MM}=0.042 vs. MAD_{BOT}=0.046, $Rank_{MM}$=0.776 vs. $Rank_{BOT}$=0.755) are better. However, for the acceptance of a regulating system, not only the absolute deviation is important, but also the deviation itself. Accordingly, a negative sign indicates, that the considered method, on average, overestimates the true efficiencies. As expected, the BOT is the method that overestimates the most, followed by the SFA. MM and DEA both underestimate.

	MAD				**MD**				**Rank**			
Method	Mean	SD	Min	Max	Mean	SD	Min	Max	Mean	SD	Min	Max
SFA	0.043	0.015	0.014	0.105	-0.006	0.024	-0.047	0.070	0.792	0.168	0.215	0.985
DEA	0.053	0.024	0.024	0.178	0.019	0.037	-0.063	0.156	0.739	0.169	0.171	0.921
BOT	0.046	0.015	0.022	0.105	-0.017	0.025	-0.080	0.058	0.755	0.166	0.194	0.924
MM	0.042	0.017	0.019	0.133	0.007	0.028	-0.055	0.113	0.776	0.171	0.186	0.974

Table 16: Performance criteria for SFA, DEA, BOT and MM. BOT: best of two method, MM: mean method.

As the averaged performance differences between the four methods (measured by MAD) are relatively small, one might expect that, for each setting, the order of advantageousness would change. In a final step, we therefore determine the "best" ("worst") method for each setting, by choosing the method with the smallest (highest) MAD (see Table 17). Thus, the MM is the best method in 54 of the 76 settings, followed by SFA (12 settings), BOT (10 settings) and DEA (0 settings). As stated above, the settings generally favor the SFA, so that the performance of the MM should improve further if this mismatch is reduced. It is also noteworthy that MM is not the poorest method for any of the settings.

	SFA	DEA	BOT	MM	*Sum*
Best Method	12	0	10	54	*76*
Percentage	15.79%	0.00%	13.16%	71.05%	*100.00%*
Worst Method	5	42	29	0	*76*
Percentage	6.58%	55.26%	38.16%	0%	*100.00%*

Table 17: Comparison of approaches for determining individual efficiency scores. BOT: best of two method, MM: mean method.

5 Conclusions

In this simulation study, we have analyzed approaches for determining individual efficiency scores, by using the two most popular estimation methods, the DEA and the SFA. Our first research objective was a systematic comparison of the two methods, using cross sectional data. Accordingly, we identified the influence factors on the performance of the particular method. We now briefly highlight the most important contributions to the literature:

1. In contrast to the literature, our results suggest that SFA can be applied to small sample sizes.

2. The percentage of DMUs on the efficient frontier influence the performance of both methods, but especially the SFA is affected.

3. We demonstrate that collinearity between the inputs has no impact on DEA.

4. The distribution of the inputs has an influence on the performance of DEA. Our attempts to shed further light on the causes of the performance variation, suggest that skewness and the relation of standard deviation and mean are the factors of influence.

5. The standard deviation of the composed error term has diverse influences on both methods.

6. The misspecification of the distribution of the inefficiency term has a crucial impact on the performance of the SFA.

7. Our results suggest that, in the majority of cases, the misspecification of the production function does not substantially affect either the SFA or the DEA.

Furthermore, we show that the different performance criteria lead in certain circumstances to diverging conclusions and that some factors have contradictory influences on the different criteria (see, for example, the variation of the standard deviation

of the composed error term). Therefore, it is particularly important to consider the appropriate criterion for both research and policy purposes. If researchers or policy makers are faced with the challenge of determining individual efficiency objectives, the mean absolute deviation (MAD) should be prioritized. Further research could confirm and extend the present results. In particular, the analysis of the influence of the input distribution on the DEA should be extended.

Due to the fact that none of the methods is absolutely superior, the combination of estimates of both methods is considered as best-practice in real-world application. Despite the fact that this procedure is also suggested in the efficiency analysis literature, it has not been analyzed in simulation studies before. Hence, we used the estimates of the first investigation step to compare two simple combination approaches with the original DEA and SFA estimates. We thereby demonstrate that the simple mean of the two methods is a compromise, which outperforms the estimates of both methods. Further research should consider how this simple approach performs in comparison to other approaches, which combine the advantages of parametric and non-parametric methods. For instance, Behr (2010) shows that the quantile regression approach can be regarded as a simple alternative for obtaining robust efficiency scores. Furthermore, a comparison with the sophisticated semiparametric frontier model, the stochastic non-smooth envelopment of data (StoNED) introduced by Kuosmanen and Kortelainen (2010), could be of interest.

References

Adler, N. and Yazhemsky, E. (2010). Improving discrimination in data envelopment analysis: PCA-DEA or variable reduction. *European Journal of Operational Research*, 202(1):273–284.

Aigner, D. and Chu, S. (1968). On estimating the industry production function. *The American Economic Review*, 58(4):826–839.

Aigner, D., Lovell, C., and Schmidt, P. (1977). Formulation and estimation of stochastic frontier production models. *Journal of Econometrics*, 6:21–37.

Andor, M. (2009). Die Bestimmung von individuellen Effizienzvorgaben - Alternativen zum Best-of-Four-Verfahren. *Zeitschrift für Energiewirtschaft*, 3:195–204.

Banker, R., Cooper, W., Grifell-Tajt, E., Pastor, J., Wilson, P., Ley, E., and Lovell, C. (1994). Validation and generalization of dea and its uses. *TOP*, 2:249–314.

Banker, R., Gadh, V., and Gorr, W. (1993). A Monte Carlo comparison of two production frontier estimation methods: Corrected ordinary least squares and data envelopment analysis. *European Journal of Operational Research*, 67(3):332–343.

Banker, R. D., Charnes, A., and Cooper, W. W. (1984). Some models for estimating technical and scale inefficiencies in data envelopment analysis. *Management Science*, 30(9):1078–1092.

Behr, A. (2010). Quantile regression for robust bank efficiency score estimation. *European Journal of Operational Research*, 200(2):568–581.

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

Cordero, J., Pedraja, F., and Daniel Santin, D. (2009). Alternative approaches to include exogenous variables in dea measures: A comparison using monte carlo. *Computers & Operations Research*, 36(10):2699 – 2706.

Farrell, M. J. (1957). The measurement of productive efficiency. *Journal of the Royal Statistical Society. Series A (General)*, 120(3):253–290.

Gong, B. and Sickles, R. (1992). Finite sample evidence on the performance of stochastic frontiers and data envelopment analysis using panel data. *Journal of Econometrics*, 51:259–284.

Haney, A. and Pollitt, M. (2009). Efficiency analysis of energy networks: An international survey of regulators. *Energy Policy*, 37(12):5814–5830.

Jensen, U. (2005). Misspecification preferred: The sensitivity of inefficiency rankings. *Journal of Productivity Analysis*, 23:223–244.

Kuosmanen, T. and Kortelainen, M. (2010). Stochastic non-smooth envelopment of data: semi-parametric frontier estimation subject to shape constraints. *Journal of Productivity Analysis*, pages 1–18.

Meeusen, W. and Broeck, J. v. D. (1977). Efficiency estimation from cobb-douglas production functions with composed error. *International Economic Review*, 18(2):435–444.

Mortimer, D. (2002). Competing Methods for Effciency Measurement: A systematic Review of Direct DEA vs SFA/DFA Comparisons. Centre for Health Program Evaluation (CHPE), Working Paper 136.

Resti, A. (2000). Efficiency measurement for multi-product industries: A comparison of classic and recent techniques based on simulated data. *European Journal of Operational Research*, 121(3):559 – 578.

Ruggiero, J. (1999). Efficiency estimation and error decomposition in the stochastic frontier model: A monte carlo analysis. *European Journal of Operational Research*, 115(3):555 – 563.

Ruggiero, J. (2007). A comparison of DEA and the stochastic frontier model using panel data. *International Transactions in Operational Research*, 14(3):259–266.

Yu, C. (1998). The effects of exogenous variables in efficiency measurement–a monte carlo study. *European Journal of Operational Research*, 105(3):569 – 580.