DIW BERLIN

1571

Discussion Papers

Deutsches Institut für Wirtschaftsforschung 2016

Semi-Parametric Measures of Scale Characteristics of German Natural Gas-Fired Electricity Generation

Stefan Seifert

IMPRESSUM

DIW Berlin
German Institute for Economic Research
Mohrenstr. 58
10117 Berlin

Tel. +49 (30) 897 89-0
Fax +49 (30) 897 89-200
http://www.diw.de

ISSN electronic edition 1619-4535

Papers can be downloaded free of charge from the DIW Berlin website:
http://www.diw.de/discussionpapers

Discussion Papers of DIW Berlin are indexed in RePEc and SSRN:
http://ideas.repec.org/s/diw/diwwpp.html
http://www.ssrn.com/link/DIW-Berlin-German-Inst-Econ-Res.html

Semi-parametric Measures of Scale Characteristics of German Natural Gas-fired Electricity Generation

Stefan Seifert*

April 13, 2016

Abstract

Scale characteristics are key properties of production functions that determine optimal firm sizes, and have considerable policy implications for sectors undergoing restructuring. However, estimates of scale characteristics typically vary with the assumptions of the underlying empirical model. This paper derives estimators of scale efficiency and scale elasticity for semi-parametric stochastic non-smooth envelopment of data (StoNED) that are based on few assumptions and rely neither on a functional form nor on distributional assumptions, but satisfy basic microeconomic properties. The estimators are applied to a unique sample covering 124 natural gas-fired power plants operating in Germany in 2011. Results indicate that on average plants operate under constant to slightly decreasing returns-to-scale, and scale inefficiency is found to be overall rather low. However, considerable improvement potential exists due to technical inefficiency. The results allow the strong fragmentation of gas-fired electricity generation in Germany, but emphasize the importance of using best practices on plant level.

JEL-Codes: D24,C14,O13,L94

Keywords: Stochastic Non-Smooth Envelopment of Data (StoNED), Returns-to-scale, Scale Elasticities, Scale Efficiency, Gas-fired Electricity Generation, Germany

*DIW Berlin – German Institute for Economic Research, Mohrenstrasse 58, D-10117 Berlin, Germany. Tel.: +49-30-89789-512, fax: +49-30-89789-200, mail: sseifert@diw.de

1 Introduction

The scale characteristics of a production function determine the optimal firm size and the market structure that uses the least resources, and, thus, can have considerable policy implications for sectors undergoing restructuring (Førsund and Hjalmarsson, 2004b). The production function and frontier estimation literature developed a variety of approaches to analyze production functions and their scale characteristics. The most prominent approaches are the non-parametric Data Envelopment Analysis (DEA) and the parametric Stochastic Frontier Analysis (SFA). For DEA, approaches have been developed to estimate scale efficiency and elasticity (see e.g. Førsund and Hjalmarsson, 2004a; Førsund et al., 2007; Podinovski and Førsund, 2010), although the results are prone to noise and depend on the assumptions regarding the underlying returns-to-scale (RTS). For parametric SFA, scale elasticity and efficiency are functions of the parameters of the assumed relationship of inputs and outputs (Ray, 1998), such as Cobb-Douglas or Translog. The parameter estimates, however, can vary with the functional relationship and with the distributional assumptions regarding the residuals (Kumbhakar et al., 2015). Thus, estimates of scale characteristics derived with DEA and SFA may depend on the assumptions by the researcher, which ultimately may influence policy implications (Bogetoft and Wang, 2005; Triebs et al., 2016). To overcome the limitations of DEA and SFA, a third approach known as Stochastic Non-Smooth Envelopment of Data (StoNED, Kuosmanen and Kortelainen, 2012) allows flexible estimation of production functions without an underlying functional form (similar to DEA) and stochastic treatment of inefficiency and noise (similar to SFA).[1]

Academic studies of the scale characteristics of the cost and production functions of fossil-fueled electricity generation discuss a range of approaches to measure scale effects, including reduced form regression with and without constraints for firm behavior, and parametric, non-, and semi-parametric frontier approaches.[2] Considerable empirical evidence exists for the US and has driven a long debate about the optimal firm and plant sizes, as well as whether minimum scales allow competitive markets (Førsund and Hjalmarsson, 2004b). For instance, Cowing and Smith (1978), who review early

[1] Several semi- and non-parametric models try to relax the assumptions of DEA and SFA, e.g., by estimating stochastic models without functional form assumptions as proposed by Fan et al. (1996). Parmeter et al. (2014) and Olesen and Petersen (2015) summarize the literature on non-parametric stochastic frontiers and stochastic DEA. Henderson and Parmeter (2009) survey the literature on constrained non-parametric regression to impose theory-driven, microeconomic conditions, such as concavity of a production function estimate.

[2] Kamerschen and Thompson Jr (1993) find that the cost characteristics of nuclear and fossil fuel steam generation differ substantially. Therefore, a separate parallel strand of the literature focuses on scale characteristics in nuclear generation. Results by Krautmann and Solow (1988) for the United States and Nemoto et al. (1993) for Japan indicate diseconomies of scale in the long run. Contrary, Arocena et al. (2012) find that larger market size benefit nuclear generation.

econometric studies of electricity generation, note that small power plants typically show increasing returns-to-scale (IRS), which, however, diminish with plant size (see e.g. Dhrymes and Kurz, 1964; Nerlove, 1963; Petersen, 1975; Christensen and Greene, 1976). Confirming this finding, Betancourt and Edwards (1987) and Maloney (2001) compare different model specifications in standard regression settings, Kopp and Smith (1980) and Goto and Tsutsui (2008) account for potential inefficiency using SFA, and Hisnanick and Kymn (1999) analyze the interplay of technical change and RTS in productivity growth. Huettner and Landon (1978) and Schmalensee and Joskow (1986) instead argue that increasing unit size reduces reliability, leading in reverse to a smaller optimal unit size with fewer outages. In support Färe et al. (1985) find IRS and decreasing returns-to-scale (DRS), depending on firm ownership. Similarly, Atkinson and Halvorsen (1984) and Huettner and Landon (1978) obtain both IRS and DRS, whereas Sueyoshi and Goto (2013) and Kumbhakar and Tsionas (2016) only find DRS. Outside the United States, there is limited empirical evidence of the scale characteristics of electricity generation. Ghosh and Kathuria (2016) indicate strong positive scale effects for India's coal-fired generation, and Akkemik (2009) finds considerable scale economies in the Turkish electricity generating sector. Based on estimates using DEA and StoNED Seifert et al. (2016) and Seifert (2015) indicate that losses due to suboptimal scale size in Germany's electricity generation depend on fuel sources. Overall, the early literature indicates IRS in electricity generation, while more recent studies suggest constant returns-to-scale (CRS) or DRS, supporting the argument that perceived RTS may have been exploited.

Unlike the other fossil fuels used for electricity generation, natural gas-fired power plants have lower capital costs and CO_2 emissions, shorter construction times, and higher operational efficiency. Their flexibility, or "rapid response" to changes in load, makes natural gas-fired plants superior for backup generation provision for intermittent wind and solar. However, to use these advantages, a minimum load of around 40 to 50% is necessary making an optimized plant size necessary (Brauner et al., 2012). Internationally, the IEA forecasts a steady increase in natural gas-fired power plant construction exceeding capacity investments in other fossil fuels (IEA, 2015, 2014).

Strong fragmentation and large differences in plant size characterize Germany's present gas-fired electricity generation sector. Hence, analysis of scale characteristics and optimal plant sizes can provide insights into the sector's transformation processes, describe the production technology more thoroughly, and help to design an optimal market structure. Motivated by the need to improve the analysis of Germany's natural gas sector, this paper proposes estimators of scale efficiency and scale elasticity based on StoNED. To incorporate the properties of the StoNED estimator, the proposed

measures account for stochasticity and the piece-wise linear shape of the estimated production function in order to satisfy the microeconomic assumptions on production technology and scale measures. The proposed estimators do not depend on any distributional or functional form assumptions, but are available with rather mild restrictions on the shape of a production function. A unique dataset of 124 natural gas-fired power plants operating in Germany in 2011 is used to test the validity of the proposed estimators for analyzing the sector's scale characteristics.
The empirical results are consistent with the recent literature. They identify significant improvement potential mainly due to technical inefficiency rather than scale inefficiency. Scale elasticity estimates indicate that most of the plants operate under constant or slightly decreasing RTS. The results infer that efficiency gains are available, but need to be realized at a plant level, and that the sector's fragmentation seems to be of minor importance from a technical perspective.
The remainder of this paper is organized as follows. Section 2 introduces StoNED and proposes new estimators of scale efficiency and scale elasticity for it. Section 3 describes the empirical set-up and Germany's gas-fired electricity generating sector. Section 4 presents the results, and section 5 concludes.

2 Methodology

Scale characteristics describe the properties of a production technology (for an overview of microeconomic characteristics of production technology see e.g. Mas-Colell et al., 1995; Färe et al., 1994). To define the technology, assume that $I (i = 1, ..., I)$ decision making units (DMUs) are observed with input-output combinations (x_i, y_i). An M-dimensional input vector x_i $(x = x_{i1}, ..., x_{iM}; x \in \mathbb{R}^M_+)$ is used to produce scalar output y_i $(y \in \mathbb{R}_+)$. A production possibility set T containing all feasible input output combinations can be written as $T = \{(x, y)|x \, can \, produce \, y\}$, with $T \subset \mathbb{R}^{M+1}_+$. The upper boundary of T is the transformation frontier F, such that $T = \{(x, y)|F(x, y) \leq 0\}$. F represents efficient input-output combinations, i.e., the points that deliver maximum output for a given level of input. Conversely, T contains all combinations, including inefficient and dominated production plans.
Next, it is assumed that T is non-empty and closed $(F \subset T)$, there is no free lunch $(x = 0 \Rightarrow y = 0)$, and inaction is possible $((x = 0, y = 0) \in T)$. Further, free disposability of inputs and outputs is given (for $x' \geq x, y' \leq y$, if $(x, y) \in T \Rightarrow (x', y') \in T$), and additivity holds (if $(x, y) \in T, (x', y') \in T \Rightarrow (x + x', y + y') \in T$). Different scaling assumptions are used to characterize the shape of the technology: constant returns-to-scale (CRS) allows arbitrary up- and downscaling of any feasible produc-

tion plan ($\forall\gamma : (x,y) \in T \Rightarrow (\gamma x, \gamma y) \in T$); non-decreasing returns-to-scale (NDRS, also called increasing returns-to-scale, IRS) allows arbitrary upscaling of any feasible production plan ($\gamma \geq 1 : (x,y) \in T \Rightarrow (\gamma x, \gamma y) \in T$); and non-increasing returns-to-scale (NDRS, also called decreasing returns-to-scale, DRS) allows only downscaling ($\gamma \in [0,1] : (x,y) \in T \Rightarrow (\gamma x, \gamma y) \in T$).[3] Finally it is assumed that T is convex (for $(x,y) \in T, (x',y') \in T, \gamma \in [0,1] \Rightarrow (\gamma x + (1-\gamma)x', \gamma y + (1-\gamma)y') \in T$), and, thus, the upper boundary of T is a concave function.[4]

2.1 Frontier and efficiency estimation with stochastic non-smooth envelopment of data (StoNED)

Scale characteristics describe the properties of a production technology that is unknown to the researcher and needs to be estimated. Since the accuracy of scale measures depends on the accuracy of the frontier estimator, several methodologies have been proposed to estimate transformation frontiers. Stochastic non-smooth envelopment of data (StoNED) as proposed by Kuosmanen (2008) and Kuosmanen and Kortelainen (2012) is a semi-parametric and stochastic approach combining characteristics of parametric SFA (Aigner et al., 1977; Meeusen and van den Broeck, 1977), and non-parametric DEA (Charnes et al., 1978; Banker et al., 1984). Similar to SFA, the approach differentiates noise and inefficiency to explain deviations from the estimated frontier based on distributional assumptions, and similar to DEA, the estimated transformation function has a piece-wise linear shape without any assumptions on an underlying functional form.

Kuosmanen and Kortelainen (2012) suggest a two stage approach to estimate a transformation frontier $y = f(x)exp(v-u)$, where $f(x)$ is the function to estimate, v is a random two-sided disturbance, and u is positive inefficiency. In the first stage, an average function $g(x)$ is estimated based on a quadratic programming problem (QP) to solve a convex non-parametric least squares problem (CNLS, Hildreth, 1954). In the second stage, an estimate of the frontier $f(x)$ is obtained by shifting $g(x)$ upwards by the expected value of inefficiency $\mu = E[u]$, which is estimated based on the distributional assumptions on v and u.

For the first stage, Kuosmanen (2008) derives a representation of the infinitely many monotonically increasing, concave, and continuous (not necessarily differentiable) functions that solve the corresponding least squares problem. Kuosmanen and Korte-

[3] In the empirical literature, variable returns-to-scale (VRS) is widely used, see e.g., Banker et al. (1984). Under VRS, the upper boundary of T combines NDRS, NIRS, and CRS characteristics.

[4] The empirical literature also considers non-convex production technologies based on the free disposability assumption, see e.g., Deprins et al. (1984) and Cazals et al. (2002).

lainen (2012) extend the approach to a production function with a multiplicative error term $\varepsilon_i = v_i - u_i$ with noise v_i and inefficiency u_i such that $y_i = f(x_i) * exp(\varepsilon_i) = f(x_i) * exp(v_i - u_i)$.[5] To estimate the average production function $g(x)$, Kuosmanen and Kortelainen (2012) derive the following non-linear QP to obtain the intercept and slope estimates based on the log-transformed multiplicative model

$$\begin{aligned} \min_{\alpha,\beta,\hat{y}} \ & \sum_{i=1}^{n} (\ln y_i - \ln \hat{y}_i)^2 \\ \hat{y}_i &= \alpha_i + \beta_i' x_i \\ \alpha_i + \beta_i' x_i &\leq \alpha_h + \beta_h' x_i \qquad \forall i, h = 1, ..., I \\ \beta_i &\geq 0 \qquad \forall i = 1, ..., I \end{aligned} \tag{1}$$

where x_i and y_i represent all observed input-output combinations. The QP tries to find the α and β coefficients minimizing the sum of the squared residuals η_i with $\eta_i = \ln y_i - \ln \hat{y}_i$. α and β are the firm-specific estimates for the intercept and slope, respectively, of a hyperplane tangent to the average production function $g(x)$. Microeconomic requirements on this hyperplanes are imposed as the following three constraints: The first constraint establishes a linear form for the estimated hyperplanes, the second constraint imposes concavity of the estimated function using Afriat's theorem (Afriat, 1967), and the third constraint imposes monotonicity. As no further restrictions are imposed on the sign of α, the estimated frontier is allowed to have VRS and may therefore have sections with increasing, constant and decreasing returns-to-scale. Note that a CRS model can be imposed by setting $\alpha_i = 0$.

The QP delivers fitted values $\hat{y}_i$ on the hyperplanes defined by the αs and βs. These $\hat{y}_i$ are typically unique, whereas the αs and βs are typically non-unique (Groeneboom et al., 2001). This non-uniqueness can lead to an overestimation of the technology violating the minimum extrapolation principle (Banker et al., 1984). To avoid such a violation a linear programming problem (LP) is used to derive the lower envelope g_{min} of the fitted values:

$$\hat{g}_{min}(x) = \min_{a \in \mathbb{R}, b \in \mathbb{R}^M} \{a + b'x \mid a + b'x_i \geq \hat{y}_i \qquad \forall i = 1, ..., I\} \tag{2}$$

The solution of (2), i.e., new intercept and slope estimates a and b, describes the average production function that is the closest envelopment of the fitted values without extrapolation with the exception of convex combinations. Contrary to the initial αs and βs, the solution to the LP (2) is unique. As Kuosmanen (2008, Theorem 4.1)

[5] Additional assumptions: u_i and v_i are assumed to be independent. v_i has a symmetric distribution with finite variance σ_v^2, u_i takes only positive values and has a finite variance σ_u^2.

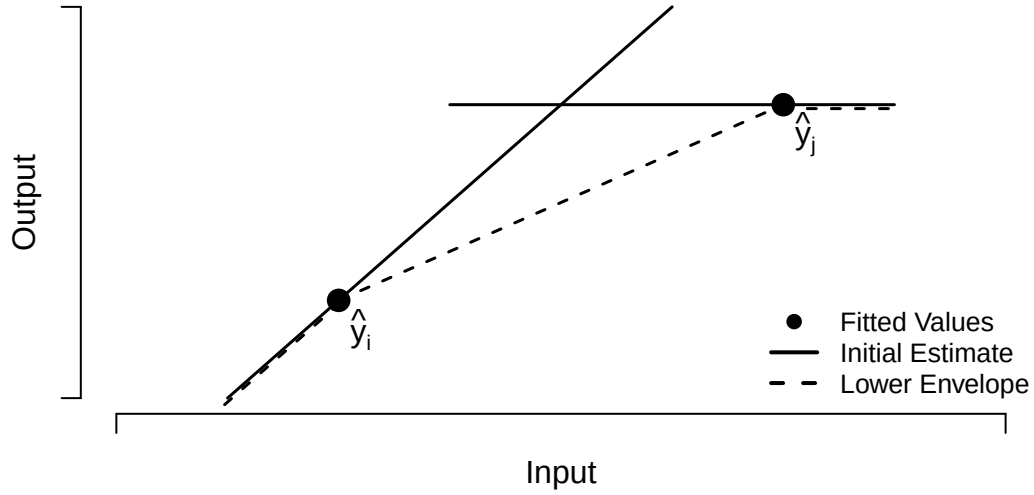

Figure 1: Illustration of Lower Envelope (Own Illustration)

shows, $\hat{g}_{min}(x)$ is identical to the DEA estimate assuming VRS for observations $(x_i, \hat{y}_i)$. Thus, $\hat{g}_{min}(x)$ has a piece-wise linear shape similar to DEA, that will be carried over later to a piece-wise linear shape of the frontier. Figure 1 illustrates this issue using an example with two observations and their fitted values i and j. The two separate hyperplanes with different intercept and slope coefficients (solid black lines), deliver the solution with minimized residuals. Recall, however, the initial estimate extrapolates the technology by more than only convex combinations of the two observed points. Solving equation (2) cures this problem by constructing the lower envelope, which is the direct connection of the fitted values (dotted black line). Figure 1 also shows that the intercept and the slope coefficients can change, which will have a direct impact on the estimated scale characteristics of the technology (see sections 2.2 and 2.3).

For the second stage, the residuals η_i are used to recover the estimates of the parameters of the distributions of inefficiency and noise, and subsequently the expected value of inefficiency.[6] Based on the estimates, $\hat{g}_{min}(x)$ is shifted to obtain a frontier estimate, but more detailed distributional assumptions are needed to derive the parameters of the distributions. Following Kuosmanen and Kortelainen (2012), a normal distribution is imposed for the noise term, $v \sim N(0, \sigma_v^2)$. The inefficiency term is assumed to take only positive values and to follow a half-normal distribution, $u \sim |N(0, \sigma_u^2)|$. Thus, the composed error term $\varepsilon_i = v_i - u_i$ is assumed to follow a normal-half-normal distribution. To recover the variance parameters σ_u and σ_v, Kuosmanen and Kortelainen (2012) suggest decomposing the residuals η_i using the pseudolikelihood estimator (PSL) proposed by

[6]Note that the fitted values from (1) are typically unique. Therefore, the residuals do not need to be recalculated against $\hat{g}_{min}(x)$.

(Fan et al., 1996, FLW).[7] For the normal-half-normal model a log-likelihood function can be expressed as a function of a single parameter $\lambda \equiv \sigma_u/\sigma_v$, with Φ denoting the cumulative distribution function of a standard normal, such that

$$\ln L(\lambda) = -n \ln \hat{\sigma} + \sum_{i=1}^{n} \ln \Phi \left[\frac{-\hat{\epsilon}_i \lambda}{\hat{\sigma}} \right] - \frac{1}{2\hat{\sigma}^2} \sum_{i=1}^{n} \hat{\epsilon_i}^2 \tag{3}$$

$$\text{with} \qquad \hat{\epsilon}_i = \hat{\eta}_i - (\sqrt{2}\lambda\hat{\sigma})/[\pi(1+\lambda^2)]^{1/2} \tag{4}$$

$$\text{and} \qquad \hat{\sigma} = \left(\left[\frac{1}{n} \sum_{i=1}^{n} \hat{\eta}_i \right] / \left[1 - \frac{2\lambda^2}{\pi(1+\lambda^2)} \right] \right)^{1/2} \tag{5}$$

Maximization of the likelihood function delivers estimates of λ, $\hat{\sigma}$, and $\hat{\epsilon}_i$. Further, $\hat{\sigma}_u = \hat{\sigma}\hat{\lambda}/(1+\hat{\lambda})$ and $\sigma_v = \hat{\sigma}/(1+\hat{\lambda})$ provide the estimates of $\hat{\sigma}_u$ and $\hat{\sigma}_v$. Given this estimate of the variance of the inefficiency, the expected value of inefficiency, $\hat{\mu}$, is then calculated as $E(u_i) = \hat{\mu} = \hat{\sigma}_u \times \sqrt{2/\pi}$. To obtain an estimate of the production function $f(x)$, the average function is shifted upwards by the corresponding expected value of inefficiency such that $\hat{f}(x) = \hat{g}_{min}(x) * exp(\hat{\mu})$. Similarly, the corresponding frontier reference points are the fitted values $\hat{y}_i$ on the average function shifted by the expected value of inefficiency, such that $\hat{y}_i^{frontier} = \hat{y}_i * exp(\hat{\mu})$.
This production function estimate has several notable characteristics. $\hat{f}(x)$ has a piecewise linear shape and can consist of up to I hyperplanes. Kuosmanen and Johnson (2010) argue, though, that the actual number of hyperplanes obtained by solving (1) and (2) typically is much lower. In other words, several observations can 'share' one hyperplane. Further, the frontier does not necessarily envelope all observations, and single observations can be outside of the estimated technology set due to the stochastic setting of the error term ε. The microeconomic assumptions on production functions outlined above are partly fulfilled. Non-emptiness of the technology is trivial as soon as data is supplied. Closedness of the technology is given and $\hat{f}(x)$ is the boundary. Free disposability and convexity of the technology is induced by the constraints in (1). Since a scaling assumption is not imposed in (1), $\hat{f}(x)$ can have VRS and may violate the additivity, the possibility of inaction, and the no free lunch assumptions.

2.2 Estimating scale efficiency

The estimated functions $\hat{g}_{min}(x)$ and $\hat{f}(x)$ have no fixed RTS assumption since the intercept a_i is unrestricted. Thus, the functions can have increasing, constant, and

[7] Kuosmanen and Kortelainen (2012) also consider a Method of Moments (MoM) estimator similar to modified ordinary least squares (MOLS). The MoM estimator is less efficient and therefore not used in this paper.

decreasing RTS along the frontier. This also means that the average product, i.e., the amount of output per input unit, can first increase, then remain constant, and finally decreases. The most productive scale size (MPSS, Banker et al., 1984) is at the input level with maximum average output and CRS. Thus, operation at this level would be beneficial for all firms.[8]

Scale efficiency can be used to quantify a firm's loss of having non-optimal size. Following Førsund and Hjalmarsson (1979), scale efficiency SE is defined as

$$SE(x_i, y_i) = \frac{\theta_i^{crs}(x_i, y_i)}{\theta_i(x_i, y_i)} \tag{6}$$

where $\theta_i(.)$ is the efficiency score against a VRS frontier, and $\theta_i^{crs}(.)$ is the efficiency score measured against a CRS frontier. Rewriting (6) into $\theta_i^{crs}(x_i, y_i) = SE(x_i, y_i) * \theta_i(x_i, y_i)$ underlines the relationship: inefficiency measured against a CRS frontier can be decomposed into the inability to use best practice defined by the VRS frontier ($\theta_i(.)$), and into the inability to use the optimal average output per input bundle ($SE(.)$). Further, it should be noted that $SE(x_i, y_i)$ measures the difference of the productivity of the MPSS and the potential productivity with current firm size as the distance between the VRS and CRS frontiers. Thus, SE is independent of a firm's managerial inefficiency since it is contained in both $\theta_i(.)$ and $\theta_i^{crs}(.)$ and cancels out.

In the microeconomic literature, it is assumed that the CRS frontier is tangent to the VRS frontier, and subsequently $T \subseteq T^{crs}$.[9] Using Shephard output distance functions (Shephard, 1970) as efficiency measures with $\theta_i^{crs}(.) \leq 1$ and $\theta_i(.) \leq 1$ it follows $\theta_i^{crs}(.) \leq \theta_i(.)$ for all i, and subsequently $SE \leq 1$. Using this definition, a scale efficiency score of one indicates optimal scale size, whereas values below one indicate productivity losses due to non-optimal size. These standard assumptions, however, my not be fulfilled in an empirical analysis based on StoNED. If a CRS frontier is additionally estimated assuming $\alpha_i = 0$ in the StoNED QP, the envelopment of the VRS by the CRS frontier is not guaranteed, but the estimated average functions or the estimated frontiers may intersect, or may not overlap. Therefore, $\theta_{StoNED}^{crs} < \theta_{StoNED}^{vrs}$ is not automatically given, and the scale efficiency estimate can be meaningless. Using the StoNED VRS estimate as a reference to fit the CRS can overcome this problem, i.e., by using an additional LP that fits the smallest enveloping cone around the frontier reference points on the

[8] In a single input model, this production plan is also the point with minimized average costs. This is not necessarily the case in a model with multiple inputs.

[9] Ray (2004) calls the CRS production function a *pseudo production function* if the true underlying RTS are VRS, since production on the CRS production function is feasible in only a few or even in only one point.

VRS estimate. This LP is constructed similar to the lower envelope in equation (2):

$$f^{crs}(x) = \underset{b^{crs}\in\mathbb{R}^M}{min}\{b'^{,crs}x \mid b'^{,crs}x_i \geq \hat{y}_i^{frontier} \qquad \forall i = 1, ..., I\} \tag{7}$$

and delivers the slope coefficients of the CRS frontier, which can consist of multiple facets. Each facet is tangent to the VRS frontier and replicates the linear up- and downscaling of the MPSS determined by the optimal output per input for different input mixes. Plugging in the observation-specific input vectors x_i delivers the fitted values on the CRS estimate, $\hat{y}_i^{crs,frontier}$, i.e., the available output for the given input if the productivity of the MPSS was reached.[10] However, note that this fitted CRS frontier relies on strictly positive inputs if outputs are produced, i.e., the data needs to satisfy the no free lunch assumption.

To measure scale efficiency in this setting, the distance between VRS and CRS frontier can be treated deterministically as the VRS frontier is consistently estimated (Kuosmanen and Kortelainen, 2012), and the CRS frontier is derived from this consistent estimate. Therefore, instead of using equation (6), a measure of scale efficiency is based on the frontier reference points and defined as

$$SE_i^{frontier} = \frac{\hat{y}_i^{frontier}}{\hat{y}_i^{crs,frontier}} \tag{8}$$

$SE_i^{frontier}$ evaluates the distance between the StoNED VRS production frontier and the fitted CRS production frontier obtained by (7). Note that this measure has the same properties as SE and allows a deterministic treatment of scale efficiency similar to DEA. Alternatively, θ_i^{crs} and θ_i in (6) can be replaced with their stochastic empirical counterparts measured against the StoNED frontier. However, stochastic estimates, such as $E[u_i|\varepsilon_i]$ proposed by Jondrow et al. (1982), are typically inconsistent in the cross-sectional setting (see e.g. Greene, 2007).

The major advantage of the proposed approach to estimate scale efficiency, i.e., its independence of any distributional assumptions, can be shown by considering an LP that fits the CRS average function $g^{crs}(x)$ around the fitted values $\hat{y}_i$ on the average function to obtain the fitted values $\hat{y}_i^{crs}$: $g^{crs}(x) = \underset{B^{crs}\in\mathbb{R}^M}{min}\{B'^{,crs}x \mid B'^{,crs}x_i \geq \hat{y}_i \qquad \forall i = 1, ..., I\}$. Compared to the LP to construct the CRS frontier, equation (7), the right-hand side is scaled up only by the expected value of inefficiency, $\hat{\mu}$. Hence, the solution vector

[10] Alternatively, one could first fit a CRS estimate around the average production function g_{min} by solving $g^{crs}(x) = \underset{b^{crs}\in\mathbb{R}^M}{min}\{b'^{,crs}x \mid b'^{,crs}x_i \geq \hat{y}_i) \qquad \forall i = 1, ..., I\}$. This estimate can then be shifted by the expected value of inefficiency $\hat{\mu}$ calculated with the VRS residuals. In most cases, shifting with an estimated inefficiency derived with the residuals against the CRS frontier will lead again to the inconsistency of VRS and CRS frontiers since they are not automatically tangent.

b is scaled up by this scalar, i.e., $b = B * exp(\hat{\mu})$. Thus, a scale efficiency measure $SE_i = \hat{y}_i/\hat{y}_i^{crs}$ for the average function is identical to the scale efficiency measure of the frontier, i.e., $SE_i = SE_i^{frontier}$. Therefore, it is irrelevant whether scale efficiency is measured against an average production function and its fitted CRS estimate, or against the frontier and its fitted CRS estimate. As a result, the scale efficiency estimate is independent of any distributional assumptions, and depends only on the assumptions imposed in equation (1), namely the concavity, monotonicity and continuity of the production function. Therefore, it could be argued that the scale efficiency measure is non-parametric, because it relies only on the non-parametric first stage of the StoNED estimate.

2.3 Estimating scale elasticities

In the context of a transformation function as outlined above, scale elasticity is a measure of the increase in output relative to a proportional increase of all inputs, evaluated as the marginal change at a point in the input-output space (Førsund et al., 2007). Assuming that a proportional increase of inputs by a factor τ leads to an increase of outputs by the factor ζ, the transformation function can be rewritten as

$$F(\tau X, \zeta(\tau, X, Y)Y) = 0 \tag{9}$$

The corresponding scale elasticity ϵ can be derived as the marginal change of the output expansion ζ caused by a marginal change in the input expansion τ over the average ratio for a differentiable function such that

$$\epsilon(X, Y) = \frac{\partial \zeta(\tau, X, Y)}{\partial \tau} \frac{\tau}{\zeta} \tag{10}$$

Using this definition, a scale elasticity greater than 1 indicates increasing returns-to-scale, a value below 1 indicates decreasing returns-to-scale, and a value equal to one indicates constant returns-to-scale. Assuming a multi-output case with N outputs $(y = y_{i1}, ..., y_{iN}; y \in \mathbb{R}_+^N)$, the derivative of (9) with respect to the input-scaling factor delivers the following rule for calculating scale elasticities (Førsund et al., 2007):

$$\frac{\partial F(\tau X, \zeta Y)}{\partial \tau} = \sum\nolimits_M \frac{\partial F(\tau X, \zeta Y)}{\partial (\tau x_m)} x_m + \sum\nolimits_N \frac{\partial F(\tau X, \zeta Y)}{\partial (\zeta y_n)} y_n \frac{\partial \zeta}{\partial \tau} = 0 \tag{11}$$

Rearranging (11) delivers the measure for scale elasticity evaluated at $\tau = \zeta = 1$

$$\begin{aligned}\epsilon(X, Y|\tau = \zeta = 1) &\equiv \frac{\partial \zeta(\tau, X, Y)}{\partial \tau} \\ &= -\sum\nolimits_M \frac{\partial F(\tau X, \zeta Y)}{\partial(\tau x_m)} x_m / \sum\nolimits_N \frac{\partial F(\tau X, \zeta Y)}{\partial(\zeta y_n)} y_n \end{aligned} \tag{12}$$

To derive scale elasticities, the transformation functions are replaced with the estimates obtained from the StoNED estimator. In the production context[11], scale elasticities are evaluated using the estimated production function $\hat{f}(x)$, which is the lower envelopment of the fitted values $\hat{y}_i$ shifted by the estimated expected inefficiency $\hat{\mu}$. Ignoring the observation specific subscripts, the function to evaluate is given by

$$\hat{f}(\tau X) = (\hat{a} + \tau \hat{b} x) * exp(\hat{\sigma_u}\sqrt{2/\pi}) = (\hat{a} + \tau \hat{b} x) * exp(\hat{\mu}) \tag{13}$$

The right-hand side in (12) for the case of multiple inputs and one output[12] y_1 and assuming $\tau = \zeta = 1$ is given by

$$\frac{\partial F(X, Y)}{\partial(x_m)} x_m = -\hat{b}_m x_m * exp(\hat{\mu}) \qquad \forall m = 1...M \tag{14}$$

$$\frac{\partial F(X, Y)}{\partial(y_1)} y_1 = (\hat{a} + \hat{b} x) * exp(\hat{\mu}) \tag{15}$$

Plugging these derivatives into (12) yields the scale elasticity measure

$$\epsilon(X, Y|\tau = \zeta = 1) = \frac{\hat{b} x * exp(\hat{\mu})}{(\hat{a} + \hat{b} x) * exp(\hat{\mu})} \tag{16}$$

that evaluates scale elasticity for a given input vector x. From (16), note that the scale elasticity depends on the estimates of a. $a > 0$ leads to $\epsilon < 1$ and indicates decreasing returns-to-scale, $a < 0$ indicates increasing returns-to-scale, and $a = 0$ indicates constant returns-to-scale. However, the a coefficient only gives the nature of the underlying RTS, whereas ϵ describes the marginal change of output given a marginal change in inputs, and delivers a more precise description of the RTS. Further, it should be noted that ϵ is independent of the estimated expected value of inefficiency $\hat{\mu}$ as it cancels out. Thus, measuring the scale elasticities on the average production function g and on the frontier f delivers identical results. This allows the estimation of the scale elasticities without any distributional assumptions since $\hat{g}_{min}$ depends only on

[11] Compare Cheng et al. (2015) for the elasticity estimation in a cost function setting

[12] For scale elasticities in production with multiple inputs and multiple outputs see Panzar and Willig (1977).

the microeconomic shape restrictions imposed in the first stage of the StoNED estimator. Similarly to the scale efficiency estimate, it can be argued that the scale elasticity measure is non-parametric, as it relies only on the non-parametric first stage of the StoNED estimate.

Recall that the two estimated functions $\hat{g}_{min}$ and $\hat{f}$ include firm-specific estimates of the intercept and slope coefficients, a_i and b_i, i.e., ϵ varies across observations due to variation in their inputs x_i and due to variation in these coefficient. Further, $\hat{g}_{min}$ and $\hat{f}$ can be non-smooth functions with a piece-wise linear shape. This may lead to non-unique solutions to equation (16) if the fitted value of observation i is a corner point. To cure this problem Banker and Thrall (1992) suggest using the right-hand side and the left-hand side partial derivatives to construct an interval of scale elasticities for the point under analysis. However, their approach must be adapted to the StoNED estimator since the fitted values for all observations can be corner points (or convex combinations of corner points).[13]

To account for the potential multiplicity of scale elasticities for an observation, scale elasticities are calculated for a set of candidate hyperplanes relevant for this observation. For observation i, the set of candidate hyperplanes C consists of the combinations of intercept and slope coefficients a and b delivering the fitted value $\hat{y}_i$ such that

$$C_i = \{(a_h, b_h) | \hat{y}_i = \hat{a}_h + \hat{b}_h x_i, h = 1, ..., I\} \tag{17}$$

The candidates can be used to construct an interval of scale elasticity measures for observation i with the boundaries ϵ_i^{min} and ϵ_i^{max}, defined as

$$\epsilon_i^{min}(X_i, y_i | (\hat{a}, \hat{b}) \in C_i) = min \ \frac{\hat{b}x * exp(\hat{\mu})}{(\hat{a} + \hat{b}x) * exp(\hat{\mu})} \tag{18}$$

$$\epsilon_i^{max}(X_i, y_i | (\hat{a}, \hat{b}) \in C_i) = max \ \frac{\hat{b}x * exp(\hat{\mu})}{(\hat{a} + \hat{b}x) * exp(\hat{\mu})} \tag{19}$$

This interval $\epsilon_i = [\epsilon_i^{min}, \epsilon_i^{max}]$ quantifies the returns-to-scale for observation i. The initial definition remains valid, and values above 1 indicate increasing returns-to-scale, values below 1 indicate decreasing returns-to-scale, and values equal to one indicate constant returns-to-scale. However, contrary to equation (16), the interval now can cover different realizations of ϵ_i, and may indicate for example IRS and DRS simul-

[13] In DEA, the problem of corner points typically occurs only for efficient units, and scale elasticities can be calculated separately for efficient and inefficient units. See Førsund et al. (2007) for a direct and an indirect approach to analyze scale elasticities in DEA.

taneously. Again, the measure is independent of the expected value of inefficiency, and therefore, of the distributional assumptions. Further, it should be noted that $\epsilon_i^{min} = \epsilon_i^{max}$ if C_i has only one element.

Taking the lower envelopment of the fitted values, enforced by (2), has a direct impact on the estimates obtained by (18) and (19). Again, Figure 1 underlines the effects: the lower envelope can influence the slope and intercept coefficients in both directions. For observation i, a second hyperplane is introduced with $a_i > \alpha_i$ and $b_i < \beta_i$, whereas for the observation j the new hyperplane is characterized by $a_j < \alpha_j$ and $b_j > \beta_j$. As a result, the scale elasticity intervals of the two observations will overlap when taking the lower envelope, while estimated elasticity will be distinctively higher for the observation i if the initial αs and βs are used.

3 Empirical set-up

3.1 Gas-fired electricity generation in Germany

Germany's electricity (and heat) generating sector is the largest in Europe and one of the largest in the world. It is highly diversified in terms of fuel sources, and while coal and lignite still predominate, renewables, especially wind and solar, play an increasing important role (BMWi, 2014). In 2014, gas-fired electricity generation accounted for about 9% of the country's total gross generation, although the share has been falling considerably with the increase of guaranteed feed-in of renewable sources. Figure 2 shows that total gas-fired generation increased steadily until 2008, and then sharply declined after 2010, while total gas-fired generation capacity still rose. Figure 3 illustrates the strong fragmentation the generation capacities: Accounting only for capacities above 10 MW, more than 50% of the 245 natural gas-fired power plants presently operating have a capacity of less than 37.5 MW, and only 49 power plants exceed a capacity of 100 MW. Figure 3 also illustrates that newly built capacities tend to be only slightly larger than existing capacities on average leading only to a slight increase of operating plants' average capacity since 1990.

From a policy perspective, two major issues on gas-fired electricity generation in Germany are currently under debate. First, gas-fired units are viewed as the technology necessary for the transition towards a carbon-free electricity system. Their operational flexibility and low CO_2 content position them as, along with pump storages, as the necessary key back-ups for intermittent renewable wind and solar. However, recent low electricity prices on the European Energy Exchange, especially due to the increasing feed-in of renewables, combined with low CO_2 certificate prices stress the profitability of gas-fired electricity generation, while the much more CO_2 intensive coal and lignite-

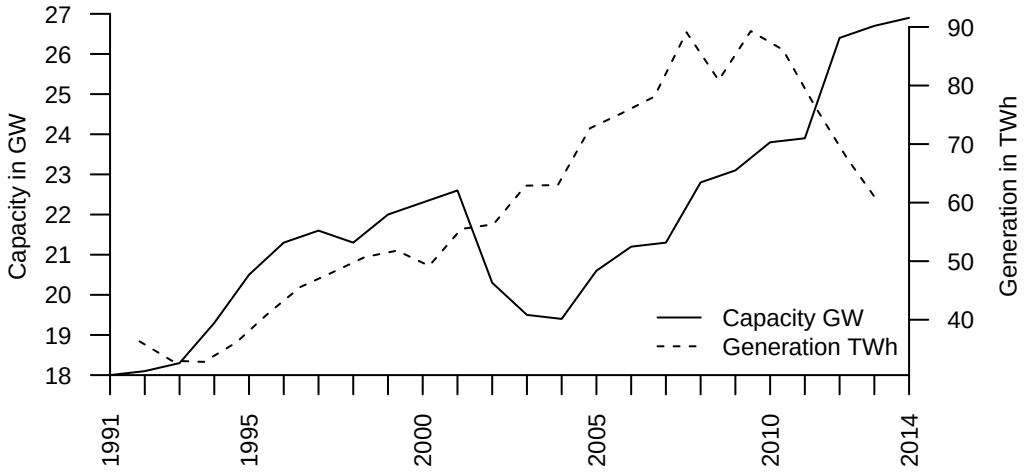

Figure 2: Gas-fired electricity generation capacity and actual generation in Germany
Own illustration, data: BMWi (2014)

based electricity generation is still profitable (Sensfuss et al., 2008). Several policy measures to allow Germany achieving its ambitious climate goals, including capacity markets and a phase-out of lignite-fired generation are under discussion (see Schrader, 2016; Reitz et al., 2014). This uncertain outlook has lead to discussion about the future of gas-fired electricity generation in Germany, and projected installation and generation varies considerably under different scenarios (compare e.g. Schlesinger et al., 2014). The second issue is the security of supply with natural gas. Concerns about disruptions of deliveries reappeared in 2014 due to the political crisis in the Ukraine and the tensions between Europe and the Russian Federation, a major source of natural gas imports. In the academic literature, natural gas delivery disruption scenarios analyzed by means of simulation models mostly indicate that a disruption would hit Eastern European countries hardest, but the price rise in Germany would be moderate (see e.g. Egging et al., 2008; Richter and Holz, 2015). Nonetheless, fears of such disruptions influence public perception of security of supply with natural gas.

3.2 Data sources, key variables and descriptive statistics

The analysis of scale characteristics of German gas-fired electricity generation uses a unique establishment level dataset for 2011 provided by the Research Data Centres of the Federal Statistical Office and the statistical offices of the Länder. The dataset is based on the monthly survey of power plants (EVAS 43311) matched with the monthly survey of the water and energy sector (EVAS 43111). For data privacy, the dataset only uses remote data processing, and detailed information, including minima and maxima

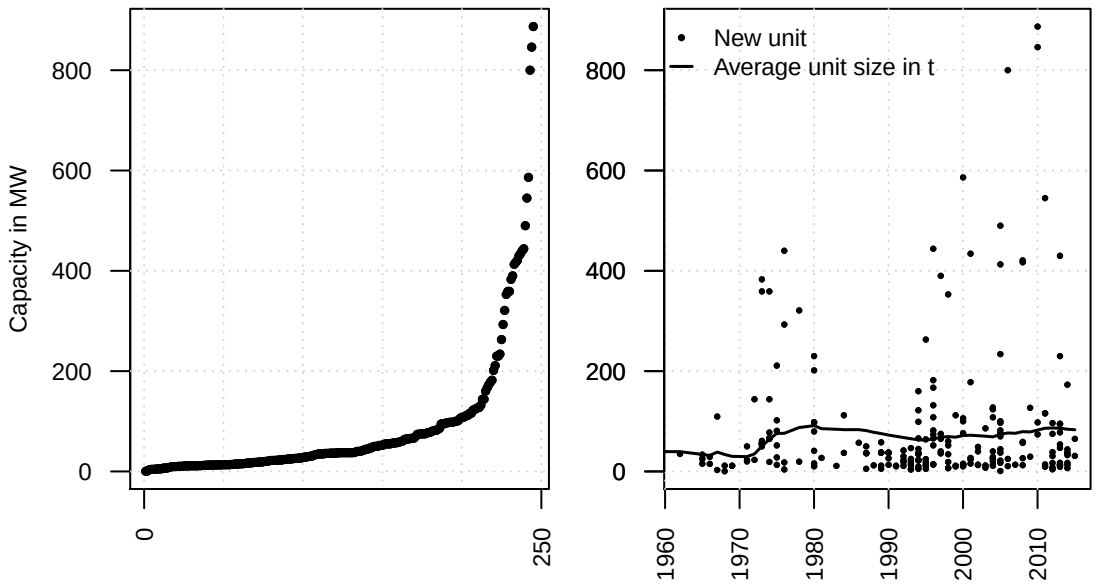

Figure 3: Gas-fired electricity generation capacity by unit size and year of construction for units exceeding 10MW
Own illustration, data: Bundesnetzagentur (2015)

of descriptive statistics and empirical results disclosing single observations, cannot be reported. Electricity generating facilities with a bottleneck capacity of at least 1 MW are included. Further, the dataset comprises large scale electricity and heat suppliers and small scale power plants for industrial use (including partial autoproducers). Observed power plants are of private, as well as of public, and mixed ownership. The sample is limited to natural gas-fired units, and other gases are excluded. The final sample consists of 124 units.

To model the production process of natural gas-fired electricity generation, a basic production model in line with other applications in the literature (e.g. Lam and Shiu, 2004; Färe et al., 1985) is used. Capital (C), labor (L), and natural gas (F) are used as inputs to produce energy (E) in the form of heat and electricity as sole output. Energy is the heat and electricity supplied, which is measured as the sum of both in GWh. On the input side, the monthly available average available capacity throughout the year in MW is used to approximate C, since this measure controls for changes in a power plant's capacity. L is measured as the sum of hours worked in 1000. F is measured using the fuel input of natural gas in TJ. Since a secondary fuel typically is used only for start-up, neglecting the secondary fuel input is expected to have minimal influence on the results. Net values are used because own consumption reduces the actual provided energy and must not influence a productivity measure. Although the model does not incorporate undesirable output directly, the natural gas input indirectly accounts for CO_2 emissions, since it is a linear transformation due to the identical CO_2 content of

	Unit	Q5	Q25	Med	Mean	Q75	Q95	SD
Capital C	MW	0.44	1.64	3.74	34.76	13.08	116.92	118.21
Fuel F	TJ	19.20	70.39	203.44	1434.92	770.42	6094.85	4052.53
Labor L	1000h	21.94	62.09	114.86	209.82	193.41	688.48	391.54
Energy E	GWh	3.93	14.99	43.34	292.87	165.31	1130.36	777.30

Table 1: Descriptive Statistics: Inputs and Outputs

natural gas in all plants.[14]

Table 1 lists the descriptive statistics of the input and output variables. The descriptive statistics show a considerable dispersion in term of plant size, and the input variables vary strongly. For C, F, and E, the 95% quantile is about 300 times higher than the 5% quantile. The dispersion is less pronounced for labor input, for which this ratio is around 30. Thus, plants with higher generation tend to be less labor-intensive than smaller plants. All plants generate both electricity and heat output.

To ensure that the sample is representative, the descriptive statistics of the final sample with 124 units are compared to the whole sample, which includes 406 units, but reduces to the final 124 units due to missing values in at least one of the variables. Corresponding comparative density plots are shown in the Figures 6 to 9 in the appendix, descriptive statistics are listed in the appendix Table 6.[15] A comparison of the two datasets indicates that the final sample represents the whole distribution of natural-gas fired power plants in Germany quite well. Although small power plants are slightly underrepresented in the matched sample, the overall distribution of power plants is well represented. The underrepresentation of very small plants should be less problematic due to the remaining number of observations in the restricted sample to define the production function in this area. Therefore, the sample is representative and allows to fully analyze scale characteristics of German gas-fired electricity generation.

4 Results

Table 2 summarizes the estimated coefficients of the lower envelope of the fitted values of the StoNED estimate, i.e., the estimated production function. Density plots of the

[14]Other emissions, as SOx and NOx, vary between the plants due to technological differences, e.g., flue gas desulphurization. Unfortunately, the technical variables are not in the dataset.

[15]A comparison of the labor input variable is not possible due to data privacy limitations of the data supplying agency.

	α_i	β_C	β_F	β_L
Min.	-6.646	0	0	0
Q25	-0.0389	0	0.1845	0
Median	0.0340	0.0547	0.2121	0.0005
Mean	36.9800	0.7714	0.1964	0.118
Q75	1.9330	0.1792	0.2146	0.0229
Max.	4256.0	23.6400	0.2417	5.078

Table 2: Summary statistics: Coefficients of $\hat{g}_{min}$

estimated coefficients are shown in Figures 10 to 13 in the appendix.[16] The explanatory power of the estimated model is high and delivers a coefficient of determination of $R^2 = 0.976$. The intercept takes positive and negative values, indicating that the estimated function consists of sections with increasing, constant and decreasing returns-to-scale. The estimated coefficients of the capital and labor input for more than 25% of the observations are zero, i.e., the marginal product of the inputs is estimated to be equal to zero, indicating that energy output is explained mainly by the fuel input. The estimated coefficient for fuel input shows little variation for the non-zero values; however, also for the fuel input variable estimated coefficients are zero for some observations.

The pseudo-likelihood estimation of the parameters of the inefficiency term delivers a variance estimate of $\sigma_u = \hat{0}.1584$, which translates into an expected inefficiency of around 12.6%. Thus, the plants could produce 12.6% additional output by using best practice with the same input endowment. This expected inefficiency deviates considerably from the results in Seifert (2015) and Seifert et al. (2016), which are both based on similar samples. However, the observation period of both studies end one in 2010, one year before the data in our sample, and one year before the considerable down-turn in gas-fired electricity generation in Germany (see Figure 2).

4.1 Scale efficiency and scale elasticity

The core of this empirical analysis is the scale characteristics of Germany's gas-fired electricity generation. The estimated scale efficiency scores are summarized in Table 3, and the corresponding density plot is shown in Figure 14 in the appendix. The overall estimated scale efficiency is rather high with an average of 98.5%, meaning that

[16] All calculations use R 3.2 (R Core Team, 2015) with the packages quadprog, alabama, bbmle and lpSolveAPI.

the losses of suboptimal plant size are on average very small. Although the minimum indicates a scale inefficiency of more than 13%, the first quantile of the distribution of scale efficiency scores indicates that only few productivity gains are available from size adjustment. The density plot reveals that scale inefficiency does not exceed 10%, except for one observation. Therefore, the expected inefficiency, indicating losses from not using best practice, clearly outweigh the scale inefficiency, indicating losses of suboptimal plant size. Therefore, firms should focus on operation efficiency, rather than the size adjustment of its power plant fleet.

	Min	Q25	Median	Mean	Q75	Max	Var
$SE_i^{frontier}$	0.8671	0.9887	0.9983	0.9854	0.9995	1	0.0007

Table 3: Summary statistics: Scale Efficiency $SE_i^{frontier}$

Summary statistics of the estimated scale elasticity intervals are listed in Table 4 and the distributions are graphically illustrated in Figure 4.[17] A considerable share of the observations' intercept estimate is close to 0, indicating close to constant returns-to-scale, yet the positive median already shows that the majority of observations in the sample operate under decreasing returns. The overall range of scale elasticity estimates is large, ranging from 0.0899 to 1.5570, and indicates a mixture of strongly increasing, constant and, strongly decreasing returns-to-scale. However, observations operate on average with a scale elasticity interval ranging from 0.9664 to 0.9944. Thus, the average gas-fired power station has slightly decreasing to constant returns-to-scale. About 25% of the observations obtain scale elasticities above 1, meaning that they operate under increasing returns-to-scale. Further, the upwards deviation from constant returns-to-scale are much lower than the downwards deviation, i.e., some plants operate with considerable DRS, while IRS are rather limited in magnitude. Figure 4 also shows that the range of the intervals decreases with the elasticities. Thus, the results indicate that corner points are especially observed for observations with lower average elasticities. Additionally, the plot indicates that the interval includes both decreasing and increasing RTS for only few observations.

From a technical perspective, the strong fragmentation of the sector with a large num-

[17] ϵ_i^{mean} and ϵ_i^{med} denote mean and median of the intervals. They are derived as mean and median of the scale elasticities calculated with the relevant candidates.

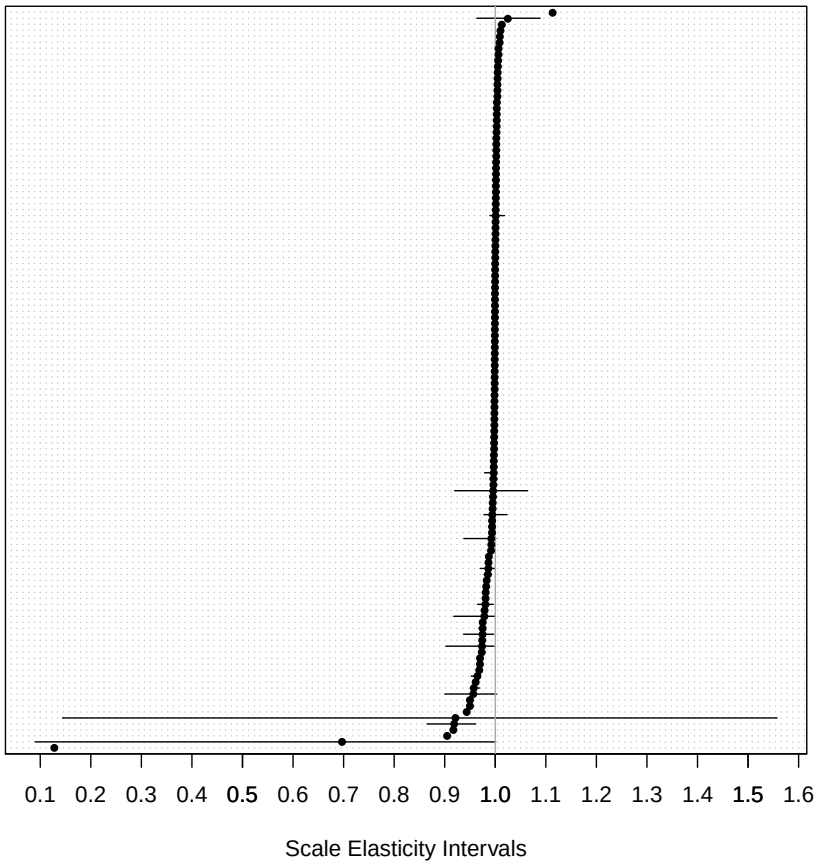

Figure 4: Mean scale elasticities and elasticity intervals

ber of small units does not pose an obstacle for highly productive gas-fired electricity generation. Indeed, if constant returns-to-scale are prevalent, the usage of smaller units allows greater flexibility due to the lower minimum load of smaller plants. Both a higher number of load hours and an higher load on average reduce wearout and increases operating life span.

The estimates of scale elasticity obtained are in line with more recent studies' findings on scale elasticities in fossil fuel electricity generation, as e.g., Kumbhakar and Tsionas (2016). Compared to Seifert et al. (2016) and Seifert (2015) also analyzing German electricity generation, results differ in a few points. Seifert et al. (2016) finds considerable scale economies that can be exploited. However, Seifert et al. (2016) analyze a panel ranging from 2003 and 2010. Their results also indicate that the best practice is determined mainly by observations from the early years of the observation period. The fact that increasing feed-in of renewables and the introduction of the European Union Emissions Trading System (EU ETS) have changed the operating environment suggests that the shape of the production functions, and, thus, the MPSS, may have

	ϵ_i^{min}	ϵ_i^{mean}	ϵ_i^{med}	ϵ_i^{max}
Min.	0.0899	0.1278	0.1278	0.1278
Q25	0.9785	0.9865	0.9915	0.9955
Median	0.9978	0.9988	0.9991	0.9997
Mean	0.9664	0.9829	0.9868	0.9944
Q75	1	1.0010	1.0010	1.0020
Max.	1.1140	1.1140	1.1140	1.5570

Table 4: Summary statistics: Scale elasticity intervals

changed over time.

The results of this analysis also need to be interpreted with caution. While the StoNED estimator allows a very flexible estimation, it is also subject to the curse of dimensionality, i.e., estimated efficiency can only increase with an increasing number of right-hand side variables. This issue can also transfer into scale efficiency scores, leading to an underestimation of scale inefficiency. Further, StoNED takes into account the existence of noise, and, as a result, the VRS and CRS estimates react less sensitively against single observations than deterministic approaches, such as DEA. Therefore, StoNED can treat a non-noisy, but extraordinarily productive observation as noisy, which can lead to an underestimation of the true technical inefficiency, scale inefficiency and scale elasticity.

4.2 Effect of the lower envelope

As outlined above, the elasticities are evaluated at the (average) production function that is derived by the lower envelope of the fitted values on this function, denoted by a_i and b_i. Alternatively, the initial estimates α_i and β_i that characterize the hyperplanes of the estimated production function could be used. Figure 5 and Table 5 compare the elasticity interval estimates for the two approaches. The results reveal that using the lower envelope has, on average, only little effect on the estimated scale elasticities, i.e., ϵ_i^{min}, ϵ_i^{mean}, ϵ_i^{med}, and ϵ_i^{max} are on average nearly identical. However, the spread of the scale elasticity estimates increases considerably when taking the lower envelopment. Further, for few observations the estimated scale elasticity decreases considerably, while the upward effect is lower in magnitude. A closer analysis of the different interval estimates indicates that larger differences can occur within one observations, e.g., with a strong decrease in ϵ_i^{min} and an increase in ϵ_i^{max}. As a result, the estimated interval ranges increase considerably by taking the lower envelope; while the initial parameter

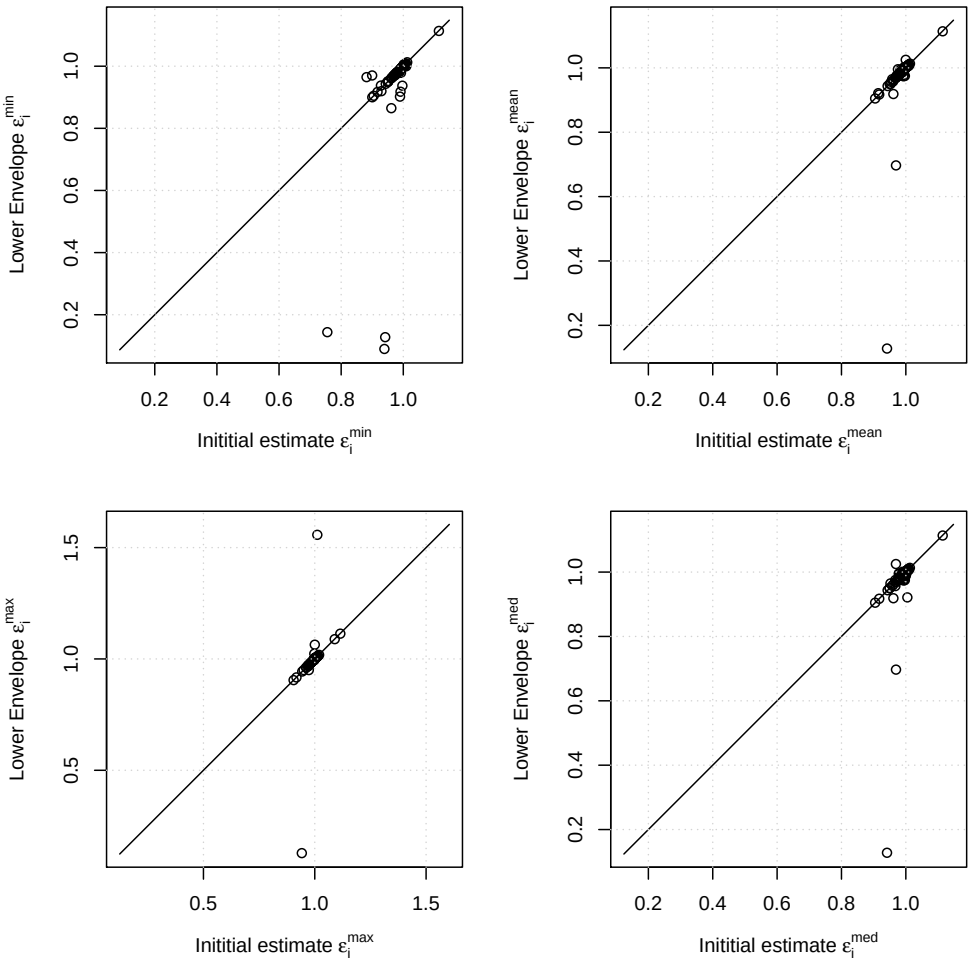

Figure 5: Comparison of elasticity intervals with and without lower envelopment

estimates result in an average interval range of 0.0097, the average range of 0.0280 obtained with the lower envelope is three times as large. Therefore, taking the lower envelope leads not only to a more cautious estimate of the technology by inducing the minimum extrapolation principle, but also to more cautious estimates of the scale elasticities because the estimated intervals are larger. Finally, the effect of taking the lower envelope might be more substantial if the shape of estimated function deviates more strongly from the CRS. This, however, can not be tested with the sample used in this paper.

	ϵ_i^{min}	ϵ_i^{mean}	ϵ_i^{med}	ϵ_i^{max}
Min.	0.7554	0.9048	0.9048	0.9048
Q25	0.9825	0.9895	0.9904	0.9961
Median	0.9981	0.9986	0.9987	0.9995
Mean	0.9864	0.9918	0.9925	0.9961
Q75	1.0010	1.0010	1.0010	1.0020
Max.	1.1150	1.1150	1.1150	1.1150

Table 5: Summary statistics: Scale elasticity intervals with intial parameter estimates

5 Conclusion

Scale efficiency and scale elasticity are the characteristics of a production function that determine optimal firm size and, thus, the market structure that uses the least resources. In the literature, several approaches to measure scale characteristics have been developed, which, however, typically rely on a number of crucial assumptions, such as the true underlying returns-to-scale, functional form assumptions, or distributional assumptions concerning deviations from the production function. The purpose of this paper is twofold: first, measures of scale elasticity and scale efficiency with less strict assumptions are derived; and, second, these new measures are used to evaluate scale characteristics of German gas-fired electricity generation for the first time.

The proposed scale efficiency and scale elasticity measures are based on stochastic non-smooth envelopment of data (StoNED). The measures take into account the stochastic nature of the two-stage StoNED approach as the estimation of a CRS production functions is adjusted to fulfill microeconomic assumptions regarding production technology. The usage of intervals of scale elasticity is proposed to account for the piece-wise linear shape of the estimated production function. It is shown that both measures are independent of distributional assumptions, do not rely on a correct returns-to-scale assumption or a functional form, allow for random noise, and are, therefore, less prone to outliers. Thus, the derived measures rely only on the consistency of the first stage of the StoNED estimator, the convex non-parametric least squares (Hildreth, 1954; Kuosmanen, 2008; Kuosmanen and Kortelainen, 2012).

Empirically, the scale characteristics of German natural gas-fired electricity generation are evaluated based on a unique dataset covering of 124 units operating in Germany in 2011. Gas-fired electricity generation is viewed as the technology that can ease the transition towards a low carbon electricity sector because of its flexibility in terms of minimum load and start-up times. However, despite these promising technical charac-

teristics, the most productive use of the inputs is only possible with optimized plant size determined by scale characteristics of the production technology. The results of the empirical analysis reveal that only few efficiency improvements are available from adjusting plants towards optimal scale size. Further, scale elasticity measures indicate that most plants already operate close to constant returns-to-scale. Thus, from a technical perspective, the results allow the strong fragmentation of Germany's natural gas-fired electricity generation. Moreover, under constant returns-to-scale, smaller units allow greater flexibility in electricity generation, reduce wearout and extend operational life. The estimates of inefficiency do, however, indicate significant technical inefficiency, i.e., differences in the use of best practice. Thus, results suggest that be realized on plant level, while the fragmentation of Germany's power plant fleet seems to be of minor importance from a technical perspective.

From a policy perspective, constant returns-to-scale of the technology are desirable as they allow a flexible design of the power plant fleet in terms of plant size. However, it is questionable whether constant returns-to-scale in the technology can be actually translated into constant economies of scale of the cost function. The current market situation, with low CO_2 and low electricity prices, favors the more CO_2 intensive baseload technologies, coal and lignite, and creates few incentives to maintain or invest in gas-fired capacities. To allow the market to support the EU's greenhouse gas emission reduction targets, investment decisions need to be based on a less uncertain future outlook, while market design needs to ensure that prices guide such allocation decisions. Several policy measures to create such incentives are under discussion (see e.g. OFGEM, 2010; Oei, 2015, for overviews), including emission performance standards (EPS), CO_2 floor prices, and capacity tenders. To support one such policy measure based on an empirical analysis, future research should consider the analysis of scale characteristics from a cost perspective.

Acknowledgements

I thank the participants of the KOMIED workshop for fruitful discussion. Further, I thank Endre Bjørndal, Astrid Cullmann, Christian von Hirschhausen, Subal Kumbhakar, Anne Neumann, Maria Nieswand, Caroline Stiel, and Michael Zschille for helpful comments, and Adam Lederer and Ann Stuart for editing.

A Appendix

	Unit	Q5	Q25	Med	Mean	Q75	Q95	SD
Capital C	MW	0.05	0.96	2.01	29.75	9.35	153.77	96.23
Fuel F	TJ	4.64	47.73	120.11	1060.59	411.45	5039.02	3230.10
Energy E	GWh	0.90	9.77	25.20	210.24	85.86	1056.32	617.77

Table 6: Descriptive Statistics: Inputs and Outputs for full sample

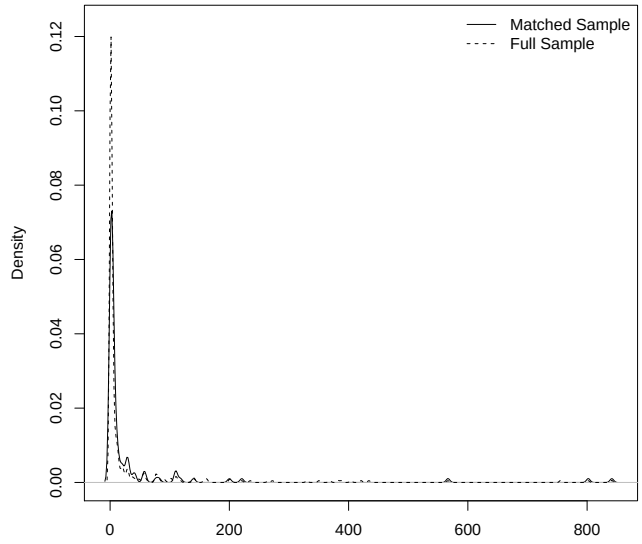

Figure 6: Density of capital input C

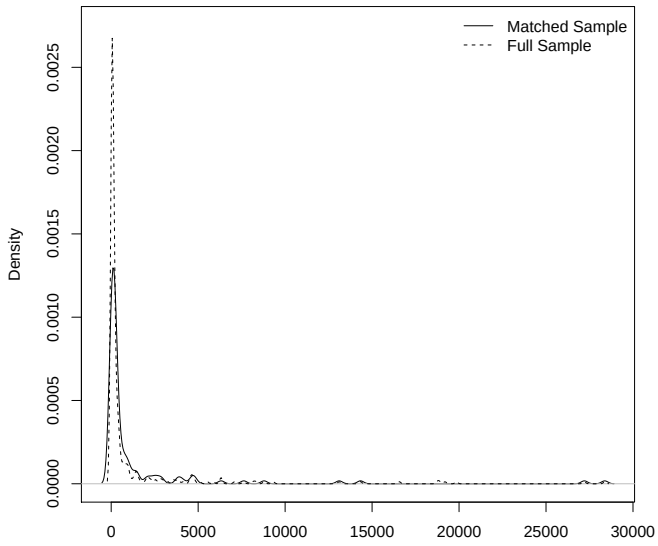

Figure 7: Density of fuel input F

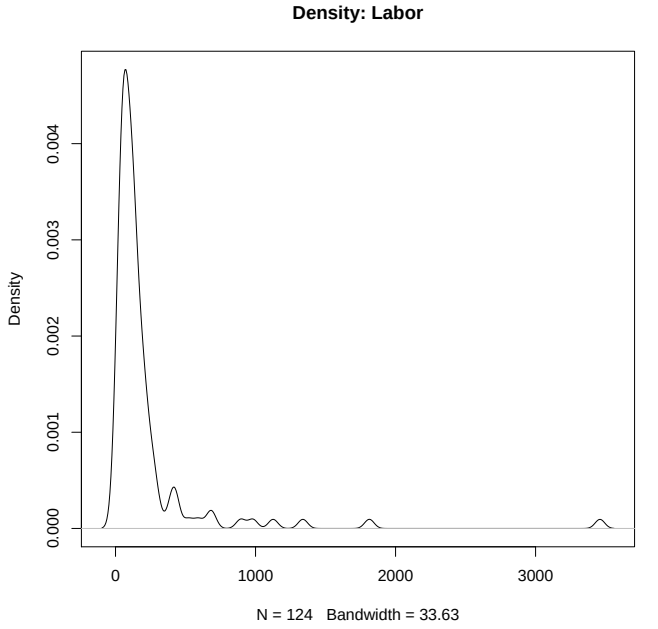

Figure 8: Density of labor input L

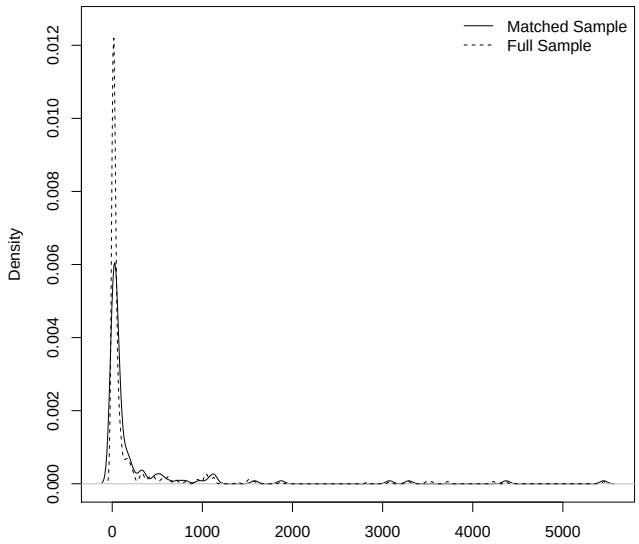

Figure 9: Density of energy output E

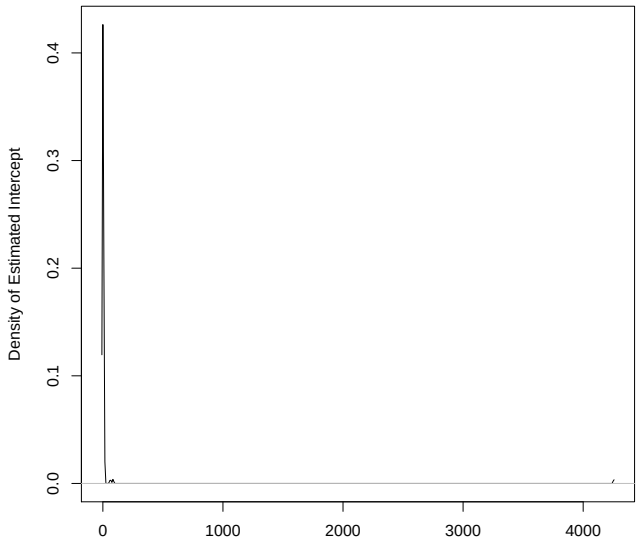

Figure 10: Density of α_i

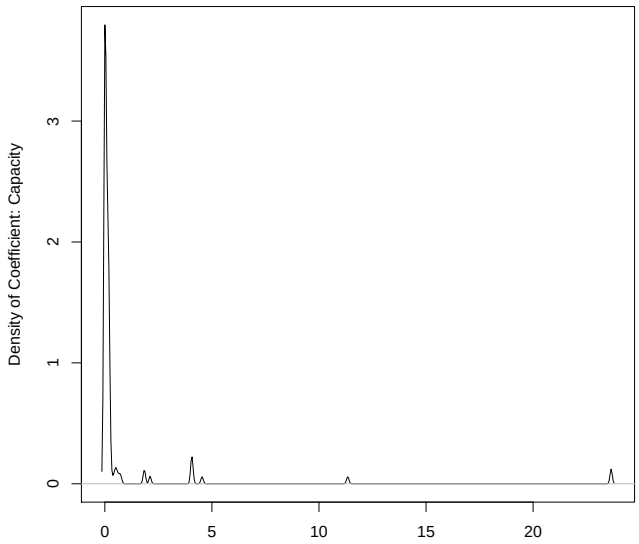

Figure 11: Density of β_C

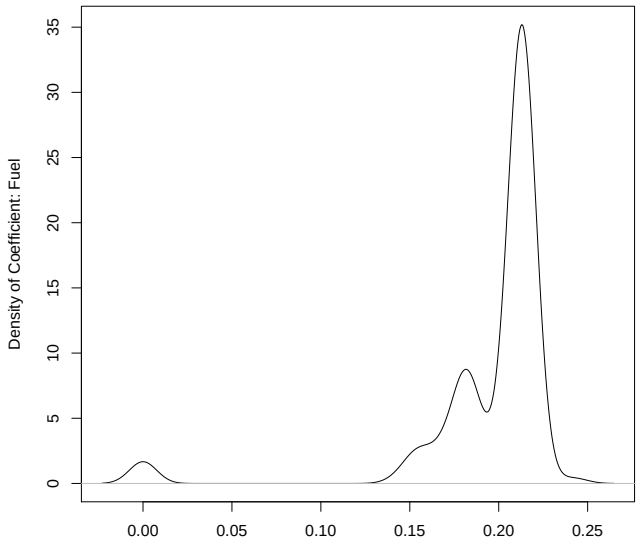

Figure 12: Density of β_F

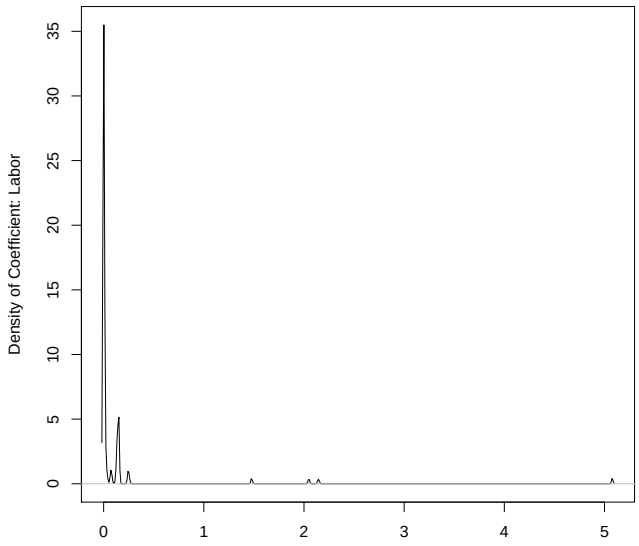

Figure 13: Density of β_L

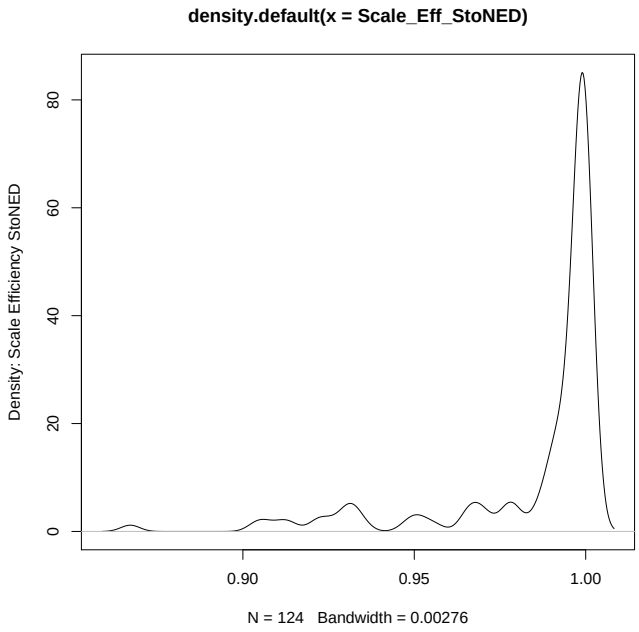

Figure 14: Density of $SE^{frontier}$

References

Afriat, S. N. (1967). The construction of utility functions from expenditure data. *International Economic Review*, 8(1):67–77.

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

Akkemik, K. A. (2009). Cost function estimates, scale economies and technological progress in the Turkish electricity generation sector. *Energy Policy*, 37(1):204–213.

Arocena, P., Saal, D. S., and Coelli, T. (2012). Vertical and horizontal scope economies in the regulated US electric power industry. *The Journal of Industrial Economics*, 60(3):434–467.

Atkinson, S. E. and Halvorsen, R. (1984). Parametric efficiency tests, economies of scale, and input demand in US electric power generation. *International Economic Review*, 25(3):647–662.

Banker, R., Charnes, R., and Cooper, W. (1984). Some models for estimating technical and scale inefficiencies in Data Envelopment Analysis. *Management Science*, 30(9):1078–1092.

Banker, R. D. and Thrall, R. (1992). Estimation of returns to scale using Data Envelopment Analysis. *European Journal of Operational Research*, 62(1):74 – 84.

Betancourt, R. R. and Edwards, J. H. Y. (1987). Economies of scale and the load factor in electricity generation. *The Review of Economics and Statistics*, 69(3):551–56.

BMWi (2014). Stromerzeugungskapazitäten, Bruttostromerzeugung und Bruttostromverbrauch. *Energiedaten, Tabelle 22.*

Bogetoft, P. and Wang, D. (2005). Estimating the potential gains from mergers. *Journal of Productivity Analysis*, 23(2):145–171.

Bundesnetzagentur (2015). Kraftwerksliste der Bundesnetzagentur. *www.bundesnetzagentur.de.*

Cazals, C., Florens, J.-P., and Simar, L. (2002). Nonparametric frontier estimation: A robust approach. *Journal of Econometrics*, 106(1):1 – 25.

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

Cheng, X., Bjørndal, E., and Bjørndal, M. (2015). Optimal scale in different environments – the case of Norwegian electricity distribution companies. *NHH Bergen Discussion Paper*.

Christensen, L. R. and Greene, W. H. (1976). Economies of scale in U.S. electric power generation. *The Journal of Political Economy*, 84(4):655–676.

Cowing, T. G. and Smith, T. G. (1978). The estimation of a production technology: A survey of econometric analyses of steam-electric generation. *Land Economics*, 54(2):156–186.

Deprins, D., Simar, L., and Tulkens, H. (1984). Measuring labor efficiency in post offices. In Marchand, M., Pestieau, P., and Tulkens, H., editors, *The Performance of Public Enterprises: Concepts and Measurements*, pages 243–267. Elsevier Science Publishing, Amsterdam North Holland.

Dhrymes, P. J. and Kurz, M. (1964). Technology and scale in electricity generation. *Econometrica: Journal of the Econometric Society*, 32(3):287–315.

Egging, R., Gabriel, S. A., Holz, F., and Zhuang, J. (2008). A complementarity model for the European natural gas market. *Energy policy*, 36(7):2385–2414.

Fan, Y., Li, Q., and Weersink, A. (1996). Semiparametric estimation of stochastic production frontier models. *Journal of Business & Economic Statistics*, 14(4):460–468.

Färe, R., Grosskopf, S., and Lovell, C. K. (1994). *Production Frontiers*. Cambridge University Press.

Førsund, F. and Hjalmarsson, L. (2004a). Calculating scale elasticity in DEA models. *Journal of the Operational Research Society*, 55(10):1023–1038.

Førsund, F. R. and Hjalmarsson, L. (1979). Generalised Farrell measures of efficiency: An application to milk processing in Swedish dairy plants. *The Economic Journal*, 89(354):294–315.

Førsund, F. R. and Hjalmarsson, L. (2004b). Are all scales optimal in DEA? Theory and empirical evidence. *Journal of Productivity Analysis*, 21(1):25–48.

Førsund, F. R., Hjalmarsson, L., Krivonozhko, V. E., and Utkin, O. B. (2007). Calculation of scale elasticities in DEA models: direct and indirect approaches. *Journal of Productivity Analysis*, 28(1-2):45–56.

Färe, R., Grosskopf, S., and Logan, J. (1985). The relative performance of publicly-owned and privately-owned electric utilities. *Journal of Public Economics*, 26(1):89 – 106.

Ghosh, R. and Kathuria, V. (2016). The effect of regulatory governance on efficiency of thermal power generation in India: A stochastic frontier analysis. *Energy Policy*, 89:11–24.

Goto, M. and Tsutsui, M. (2008). Technical efficiency and impacts of deregulation: An analysis of three functions in US electric power utilities during the period from 1992 through 2000. *Energy Economics*, 30(1):15–38.

Greene, W. H. (2007). The econometric approach to efficiency measurement. In Fried, H., Lovell, C. K., and Schmidt, S., editors, *The Measurement of Productive Efficiency*, chapter 2, pages 92–250. Oxford University Press, Oxford.

Groeneboom, P., Jongbloed, G., and Wellner, J. A. (2001). Estimation of a convex function: Characterizations and asymptotic theory. *Annals of Statistics*, 29(6):1653–1698.

Henderson, D. J. and Parmeter, C. F. (2009). Imposing economic constraints in non-parametric regression: survey, implementation, and extension. *Advances in Econometrics*, 25:433.

Hildreth, C. (1954). Point estimates of ordinates of concave functions. *Journal of the American Statistical Association*, 49(267):598–619.

Hisnanick, J. J. and Kymn, K. O. (1999). Modeling economies of scale: the case of US electric power companies. *Energy Economics*, 21(3):225–237.

Huettner, D. A. and Landon, J. H. (1978). Electric utilities: scale economies and diseconomies. *Southern Economic Journal*, 44(4):883–912.

IEA (2014). *World energy outlook*. OECD/IEA.

IEA (2015). *Key world energy statistics*. International Energy Agency.

Jondrow, J., Knox Lovell, C., Materov, I., and Schmidt, P. (1982). On the estimation of technical inefficiency in the stochastic frontier production function model. *Journal of Econometrics*, 19(2/3):233–38.

Kamerschen, D. R. and Thompson Jr, H. G. (1993). Nuclear and fossil fuel steam generation of electricity: differences and similarities. *Southern Economic Journal*, 60(1):14–27.

Kopp, R. J. and Smith, V. K. (1980). Frontier production function estimates for steam electric generation: A comparative analysis. *Southern Economic Journal*, 46(4):1049–1059.

Krautmann, A. C. and Solow, J. L. (1988). Economies of scale in nuclear power generation. *Southern Economic Journal*, 55(1):70–85.

Kumbhakar, S. C. and Tsionas, E. G. (2016). The good, the bad and the technology: Endogeneity in environmental production models. *Journal of Econometrics*, 190(2):315 – 327.

Kumbhakar, S. C., Wang, H., and Horncastle, A. P. (2015). *A Practitioner's Guide to Stochastic Frontier Analysis Using Stata*. Cambridge University Press.

Kuosmanen, T. (2008). Representation theorem for convex nonparametric least squares. *Econometrics Journal*, 11(2):308–325.

Kuosmanen, T. and Johnson, A. L. (2010). Data envelopment analysis as nonparametric least-squares regression. *Operations Research*, 58(1):149–160.

Kuosmanen, T. and Kortelainen, M. (2012). Stochastic non-smooth envelopment of data: semi-parametric frontier estimation subject to shape constraints. *Journal of Productivity Analysis*, 38(1):11–28.

Lam, P.-L. and Shiu, A. (2004). Efficiency and productivity of China's thermal power generation. *Review of Industrial Organization*, 24(1):73–93.

Maloney, M. T. (2001). Economies and diseconomies: Estimating electricity cost functions. *Review of Industrial Organization*, 19(2):165–180.

Mas-Colell, A., Whinston, M., and Green, J. (1995). *Microeconomic Theory*. Oxford student edition. Oxford University Press.

Meeusen, W. and van den Broeck, J. (1977). Efficiency estimation from Cobb-Douglas production functions with composed error. *International Economic Review*, 18(2):435–444.

Nemoto, J., Nakanishi, Y., and Madono, S. (1993). Scale economies and overcapitalization in Japanese electric utilities. *International Economic Review*, 34(2):431–440.

Nerlove, M. (1963). Returns to scale in electricity supply. In Christ, C. F., editor, *Measurement in Economics - Studies in Mathematical Economics and Econometrics in Memory of Yehuda Grunfeld*, chapter 7, pages 167–198. Stanford University Press.

Oei, P.-Y. (2015). *Decarbonizing the European Electricity Sector - Modeling and Policy Analysis for Electricity and CO_2 Infrastructure Networks*. PhD thesis, TU Berlin.

OFGEM (2010). Project discovery: Options for delivering secure and sustainable energy supplies. Technical report, Office of Gas and Electricity Markets.

Olesen, O. B. and Petersen, N. C. (2015). Stochastic Data Envelopment Analysis - A review. *European Journal of Operational Research*.

Panzar, J. C. and Willig, R. D. (1977). Economies of scale in multi-output production. *The Quarterly Journal of Economics*, 91(3):481–493.

Parmeter, C. F., Kumbhakar, S. C., et al. (2014). Efficiency Analysis: A Primer on Recent Advances. *Foundations and Trends (R) in Econometrics*, 7(3-4):191–385.

Petersen, H. C. (1975). An empirical test of regulatory effects. *The Bell Journal of Economics*, 6(1):111–126.

Podinovski, V. V. and Førsund, F. R. (2010). Differential characteristics of efficient frontiers in Data Envelopment Analysis. *Operations research*, 58(6):1743–1754.

R Core Team (2015). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Ray, S. C. (1998). Measuring scale efficiency from a translog production function. *Journal of Productivity Analysis*, 11(2):183–194.

Ray, S. C. (2004). *Data Envelopment Analysis: Theory and Techniques for Economics and Operations Research*. Cambridge University Press.

Reitz, F., Gerbaulet, C., Hirschhausen, C. v., Kemfert, C., Lorenz, C., and Oei, P.-Y. (2014). Verminderte Kohleverstromung könnte zeitnah einen relevanten Beitrag zum deutschen Klimaschutzziel leisten. *DIW Wochenbericht*, 47:1219–1229.

Richter, P. M. and Holz, F. (2015). All quiet on the eastern front? Disruption scenarios of Russian natural gas supply to europe. *Energy Policy*, 80:177–189.

Schlesinger, M., Lindenberger, D., and Lutz, C. (2014). Development of energy markets – energy reference forecast. Technical report, Project No. 57/12, Study commissioned by the German Federal Ministry of Economics and Technology.

Schmalensee, R. and Joskow, P. L. (1986). Estimated parameters as independent variables: An application to the costs of electric generating units. *Journal of Econometrics*, 31(3):275 – 305.

Schrader, C. (2016). Can Germany engineer a coal exit? *Science*, 351(6272):430–431.

Seifert, S. (2015). Measuring productivity when technologies are heterogeneous: A semi-parametric approach for electricity generation. *DIW DP 1526*.

Seifert, S., Cullmann, A., and von Hirschhausen, C. (2016). Technical efficiency and CO_2 reduction potentials - An analysis of the German electricity and heat generating sector. *Energy Economics*, 56:9 – 19.

Sensfuss, F., Ragwitz, M., and Genoese, M. (2008). The merit-order effect: A detailed analysis of the price effect of renewable electricity generation on spot market prices in Germany. *Energy policy*, 36(8):3086–3094.

Shephard, R. (1970). *Theory of Cost and Production Functions*. Number 4 in Princenton Studies in Mathematical Economics. Princenton University Press.

Sueyoshi, T. and Goto, M. (2013). Returns to scale vs. damages to scale in data envelopment analysis: An impact of U.S. clean air act on coal-fired power plants. *Omega*, 41(2):164–175.

Triebs, T. P., Saal, D. S., Arocena, P., and Kumbhakar, S. C. (2016). Estimating economies of scale and scope with flexible technology. *Journal of Productivity Analysis*, 45(2):173–186.