DIW BERLIN

1283

Discussion Papers

Deutsches Institut für Wirtschaftsforschung 2013

Estimating Alternative Technology Sets in Nonparametric Efficiency Analysis

Restriction Tests for Panel and Clustered Data

Anne Neumann, Maria Nieswand and Torben Schubert

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 print edition 1433-0210
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

Estimating Alternative Technology Sets in Nonparametric Efficiency Analysis: Restriction Tests for Panel and Clustered Data

Anne Neumann* Maria Nieswand† Torben Schubert‡

February 25, 2013

Abstract

Nonparametric efficiency analysis has become a widely applied technique to support industrial benchmarking as well as a variety of incentive-based regulation policies. In practice such exercises are often plagued by incomplete knowledge about the correct specifications of inputs and outputs. Simar and Wilson (2001) and Schubert and Simar (2011) propose restriction tests to support such specification decisions for cross-section data. However, the typical oligopolized market structure pertinent to regulation contexts often leads to low numbers of cross-section observations, rendering reliable estimation based on these tests practically unfeasible. This small-sample problem could often be avoided with the use of panel data, which would in any case require an extension of the cross-section restriction tests to handle panel data. In this paper we derive these tests. We prove the consistency of the proposed method and apply it to a sample of US natural gas transmission companies in 2003 through 2007. We find that the total quantity of gas delivered and gas delivered in peak periods measure essentially the same output. Therefore only one needs to be included. We also show that the length of mains as a measure of transportation service is non-redundant and therefore must be included.

*Universität Potsdam, Dept. of Economics and Social Science for Economic Policy, August-Bebel-Strasse 89, D-14482 Potsdam, Germany

†DIW Berlin, Mohrenstrasse 58, D-10405 Berlin, Germany

‡Fraunhofer ISI, Breslauer Strasse 48, D-76139 Karlsruhe, Germany

JEL-Codes: C14, L51, L95
Keywords: Benchmarking models, Network industries, Nonparametric efficiency estimation, Data envelopment analysis, Testing restrictions, Subsampling, Bootstrap

1 Introduction

Nonparametric efficiency analysis has become increasingly important for sound decision-making in a variety of economic research fields. In addition to industrial benchmarking, the regulation of network industries, among them natural gas transmission, is a considerable field of application. Regulatory decisions are often directly contingent on the results of such analyses, e.g. in Norway, Germany, and Austria. Because the decisions have strong financial implications for both customers and firms, it is critical that the models underlying the analyses are specified correctly. In the context of efficiency-estimation this means that the correct inputs and outputs are accounted for.

Restriction tests, proposed by Simar and Wilson (2001) and Schubert and Simar (2011), allow for the testing of hypotheses regarding the inputs and outputs. Nevertheless, the practical value of these tests is limited when the cross-section sample size is small. This is typically the case in monopolized and oligopolized producer markets. One way to solve this small sample problem is to observe firms over time, that is, to use panel data. The regular cross-section tests are then, however, not longer applicable, because of the *i.i.d.* assumption. Therefore an extension is required that allows for correlation across the time dimension. These tests are developed in the course of this paper and proven to be consistent if the production frontier is constant over time.

The paper is organized as follows: Section 2 provides an expository overview of the role of efficiency measurement and benchmarking as a regulatory tool. We explain some of the benefits of nonparametric techniques as well as major difficulties that arise from small cross-section sample sizes rendering reliable estimation often impossible. We argue that restriction tests for clustered data could help solving this problem in many contexts. In Section 3 we describe the proposed test procedures. Section 4 describes our data set and presents the results. Section 5 concludes.[1]

[1]We thank Luis Orea and the participants of the 5th International Workshop on Empirical Methods in Energy Economics (EMEE), the annual meeting of the *Verein für Socialpolitik* 2012, and the 10th Conference on Applied Infrastructure Research (INFRADAY) for valuable comments and discussions.

2 Efficiency Measurement as a Decision-Making Tool

Data Envelopment Analysis (DEA) is a nonparametric method for efficiency analysis and is closely related to the classical models of activity analysis.[2] It offers an alternative way to evaluate the performance of production entities.[3] Unlike classical activity analysis, the concept of efficiency analysis intends to express productive efficiency in a multiple-input-multiple-output framework and thereby avoids the index number problem (Farrell, 1957; Cooper et al., 2011). In efficiency analysis, the performance of a production unit is determined by comparing it to a group of production entities that have access to the same transformation process (technology) through which they convert the same type of resources (inputs) into the same type of products (outputs). From the observed input-output-combinations a best practice (frontier) is constructed against which each entity is assessed individually. The distance to that frontier reflects the production unit's ability to transform inputs into outputs, relative to what empirically is found and therefore assumed to be feasible.

Hence, efficiency analysis provides a quantitative measure of the existing potential for improvement. As pointed out by Bogetoft and Otto (2011), the scope of application of the DEA method is rich, since conceivable production entities include firms, organizations, divisions, industries, projects, decision-making units (DMUs), and individuals. Empirical analysis investigates, for example, industrial entities such as warehouses (Schefczyk, 1993) and coal mines (Thompson et al., 1995). As noted by Schefczyk (1993) industrial benchmarking serves as a tool to generate measures by which corporate decision-making can be brought in line with the corporate goal of operating efficiently. DEA in combination with Malmquist indices is also commonly applied to determine technical efficiency change, technical change, total factor productivity change, see e.g. Jamasb et al. (2008), all of which are useful tools to evaluate a particular sector and regulatory changes. In addition, DEA is widely used in the regulation of network industries in order to overcome disincentives and distortions related to monopolistic market structures.

[2] For more details about the methodological linkages of activity and efficiency analysis see Färe and Grosskopf (2005).

[3] Homburg (2001) gives detailed insights on how nonparametric efficiency analysis can contribute to activity-based management.

2.1 Benchmarking and DEA in Regulation

It is well known that the private sector draws on comparative analyses, such as activity analysis, to improve its performance. Starting in the 1990s, regulatory authorities are making increasing use of benchmarking techniques in order to facilitate incentive regulation of network utilities; see e.g., Jamasb and Pollitt (2003). In particular, electricity and natural gas transmission and distribution utilities are involved in regulatory activities; see e.g., Jamasb et al. (2004); Cullmann (2012); Farsi et al. (2007); Sickles and Streitwieser (1998); Hollas et al. (2002). Applying benchmarking methods allows the regulator to simulate competitive market structures (quasi-competition), thus helping to pursue and implement regulatory objectives, e.g., reducing monopolistic power and promoting the efficient use of resources. Beside process and activity analysis as well as parametric frontier models, e.g. Stochastic Frontier Analysis (SFA), regulators frequently rely on DEA in order to establish benchmarks for target determination (Haney and Pollitt, 2009).

Due to the market structure in network industries, many regulatory benchmarking applications rely on a small number of observations; see e.g., Jamasb et al. (2008). Larger sample sizes can generally be obtained two ways: First, using cross-country analysis, and second, using cross-sectional data across multiple time periods. When pooling observations across countries, simple cross-section tests can be used as long as we guarantee that all countries have access to the same technology. However, when comparing the same individuals across time, the additional problem emerges that a firm's present and past observations are generally not independent. So pure cross-section methods will lead to false inference, even if the technology did not change over the respective time period.

2.2 A Need for Specification Analyses

As a nonparametric method, DEA has, on the one hand, appealing characteristics (Simar and Wilson, 2008); beside its great flexibility and easy computability, it requires only few assumptions on the technology set and its frontier. Particularly, it does neither assume a distribution for the inefficiency term nor does it impose a functional form to express the production process generating the observed input-output-combinations (Haney and Pollitt, 2009; Simar and Wilson, 2008). On the other hand, the DEA estimator has drawbacks that are highly relevant for both regulatory and industrial performance analysis. In addition to its outlier sensitivity, its non-parametric nature dramatically reduces its asymptotic convergence rate

when the dimensionality of the production possibility set is high. This is particularly problematic when only a few number of observations are present and, hence, some argue that DEA is not an ideal tool for regulatory purposes; see e.g. Shuttleworth (2005). The critique also extends to cases where the analysis of total factor productivity change in regulated and non-regulated sectors is of primar interest.

Because the speed of convergence decreases with an increasing number of inputs and outputs, specifying the model economically is even more important than in parametric models. This means that we should exclude inputs and outputs that do not contribute to the model. However, in most situations it is uncertain what is the correct specification of the technology set. For example, uncertainty may occur as a result of information asymmetries, when the analyst lacks full information about the precise production process.[4] Then statistical inference about alternative specifications is desirable in order to make sound decisions about the reasonable choice of variables.

Simar and Wilson (2001, 2011) propose different restriction tests for nonparametric efficiency analysis allowing to investigate whether certain variables can be excluded (exclusion restriction) or summed up (aggregation restriction). Schubert and Simar (2011) extend these tests by introducing a subsampling procedure (a special kind of bootstrap) that relaxes the homogeneity assumption[5] and, therefore, allows for tests in input and output directions within the same dataset. Although the benefits of restriction tests on production process formulations are obvious, in the applied literature they receive only scant attention. Restriction tests notably improve nonparametric benchmarking, because they increase the confidence in the chosen representation of the production process by providing statistical inference. The risk of overestimating the performance due to the 'curse of dimensionality' is reduced when variables are identified as irrelevant and consequently are excluded from further investigation.[6] Yet the existing im-

[4]The information asymmetries in regulation mainly result from *adverse selection* and *moral hazard* problems (Joskow, 2006).

[5]The homogeneity assumption is comparable to the parametric homoscedasticity assumption and means that the distribution of the inefficiencies does not depend on inputs or the outputs. The problem is that it will not generally hold in both the input and the output direction, prohibiting tests based on it in both directions.

[6]Alternatively, variables could be omitted or aggregated. Omitting variables based on correlations should be avoided for translation invariant DEA models (Dyson et al., 2001) and aggregating variables based on principal components might be inappropriate for radial efficiency measurement (Simar and Wilson, 2001). However, the restriction tests proposed by Simar and Wilson (2001) and Schubert and Simar (2011) provide statistical inference procedures for the investigation of aggregates.

plementations of the proposed tests are restricted to cross-sectional data only and are therefore not applicable to (unbalanced) panel data.

We aim to present a test procedure that is able to account for correlation likely being present when panel (respectively clustered data) is used.[7] The contribution of the paper at hand is twofold: First, we further develop the theoretical underpinnings of the restriction tests in order to enhance their applicability to (unbalanced) panel and clustered data in general. This requires accounting for intra-observational dependencies. Second, we demonstrate the relevance of the proposed test procedure for benchmarking by applying the method to a data set of US natural gas transmission companies.[8] Clearly, the main benefits of the proposed approach are improving the efficiency estimation and overcoming lacks of information regarding the production process. Although our demonstration relates to the regulatory framework, it is straightforward to apply the technique to any other setting where the mentioned problems arise.

3 Methodology

3.1 Technology Estimation using the DEA Estimator

We start by presenting the analytical framework. It introduces the concepts necessary for the later proofs of consistency for the test statistics.

Let $x^i \in \mathbb{R}^p_+$ and $y^i \in \mathbb{R}^q_+$ denote the vectors of p inputs and q outputs. The technology set Ψ represents the feasible input-output-combinations available to firm i (Bogetoft and Otto, 2011) and can be defined as

$$\Psi = \left\{ (x, y) \in \mathbb{R}^{p+q}_+ \mid x \,\text{can produce}\, y \right\}. \tag{1}$$

For Ψ we assume free disposability and convexity. The boundary of Ψ, denoted by Ψ^δ, describes the efficient production frontier, i.e. the technology, and can be defined as

$$\Psi^\delta = \left\{ (x, y) \in T \mid \left(\gamma x, \gamma^{-1} y\right) \notin \Psi \,\text{for any}\, \gamma < 1 \right\}. \tag{2}$$

According to Equation 2, a firm that employs a production plan that belongs to Ψ^δ, is regarded as efficient and its input-output combination cannot be

[7] Note, that panel data is just one example of clustered data and that therefore, the applicability of the proposed test is even more comprehensive.

[8] This industry is subject to analysis concerned with total factor productivity growth and technical change in the light of changing regulation; see e.g. Sickles and Streitwieser (1992, 1998); Jamasb et al. (2008).

improved. Companies that operate at points in the interior of Ψ exhibit inefficiencies (Simar and Wilson, 2001), which can be diminished by moving toward the efficient frontier. Being able to handle multi-input and multi-output settings, the Debreu-Farrell measure[9] quantifies the respective firm-individual degree of efficiency. For any particular coordinate $(x_0, y_0) \in \Psi$, the Debreu-Farrell efficiency score is determined by the radial distance from (x_0, y_0) to the efficient frontier Ψ^δ. It expresses the maximal proportional contraction of all inputs x that allows to produce output level y for input-orientation, and the maximum proportional expansion of all outputs y that is feasible with the given inputs x, for output-orientation, respectively.

We restrict ourselves to the input-orientated firm-specific efficiency measure, which can formally be expressed as

$$\theta(x_0, y_0) = inf\{\theta \geq 0 \mid (\theta x_0, y_0) \in \Psi\}. \tag{3}$$

Hence, if $\theta(x_0, y_0) = 1$, the company is efficient and operates along the frontier Ψ^δ. If $\theta(x_0, y_0) \leq 1$, the company can improve its performance by reducing its input quantities proportionally. Together with the imposed assumptions, Equations 1 and 2 set up the *true* economic production model and characterize the data generating process $\mathbb{P}$ (DGP).[10] However, the *true* technology set Ψ, and hence, the *true* efficient technology Ψ^δ against which observations are compared to, are unknown and both need to be estimated from the observed input-output-combinations.

To approximate Ψ, we apply the DEA estimator proposed by Banker et al. (1984), which incorporates the assumptions of free disposability, convexity and variable returns to scales. Thus, the linear program estimating the unknown input-oriented efficiency score θ becomes:

$$\begin{aligned} \hat{\theta}(x_0, y_0) = \min_{\theta, \lambda_1, \ldots, \lambda_n} \{\theta > 0 \mid\ & \theta x_0 \geq \textstyle\sum_{i=1}^n \lambda^i x_k^i; k = 1, \ldots, p \\ & y_0 \leq \textstyle\sum_{i=1}^n \lambda^i y_l^i; l = 1, \ldots, q \\ & \textstyle\sum_{i=1}^n \lambda^i = 1;\ \lambda^i \geq 0 \,\forall\, i = 1, \ldots, n\}. \end{aligned} \tag{4}$$

It is well known that the rate of convergence for nonparametric estimators, such as DEA, is small compared to parametric estimators (Simar and Wilson, 2008). The consistency of this estimator is proven by Kneip et al.

[9] This measure is based on the work of Debreu (1951) and Farrell (1957). Alternatively, the concept proposed by Shepard (1970) can be used.

[10] To comprehensively define the DGP, assumptions on the statistical model are necessary. Due to space limitations, we omit the discussion and refer the reader to e.g., Simar and Wilson (2001).

(1998). But like most nonparametric estimators it suffers from the 'curse of dimensionality', which implies that the rate of convergence (i.e. the speed by which the estimation errors are reduced in sample size) goes down as the number of inputs and outputs increases. Additionally, the DEA estimates are upward biased. This implies that the *true* efficiency is lower than the one estimated in finite samples. The precision of the estimation results is significantly affected by the ratio of observations to the number of variables and a considerable interest arises to test for the relevance of particular inputs and outputs. Reducing the dimensionality of the technology set Ψ by possibly irrelevant variables can offer substantial gains in estimation efficiency and decrease finite sample biases.

3.2 Testing Restrictions

Having specified the estimation approach, we formulate the restrictions on the technology set that we aim to test. It is our objective to test whether particular outputs are relevant for modeling the technology set appropriately. Although we focus on the relevance of outputs in this paper, we note that the method is broader. Alternatively, the relevance of input variables can be considered. Further, it can be tested whether inputs and outputs are individually relevant contributors to production or if they can be aggregated. We extend a test procedure suggested by Simar and Wilson (2001) to panel data while, following Schubert and Simar (2011), using subsampling procedures. The formalism of proofs of consistency in the appendix is independent of whether restriction is due to an exclusion or due to an aggregation restriction.

The basic idea of the original approach is to compare efficiency estimates obtained from a technology set including all potential outputs with efficiency estimates obtained from a restricted technology set that excludes at least one output (or aggregates at least two outputs). The rationale behind assigning a particular output as possibly irrelevant is the uncertainty regarding its relationship to the considered input(s). An output is identified as redundant if the difference between the estimates of both technology sets, where the restricted is nested in the unrestricted, do not differ significantly. Conceptionally this implies that the irrelevant output is not produced by the firm or, putting it differently, the considered inputs do not contribute to the output in question. The main benefit of this approach is twofold: First, selecting outputs can be based on statistical tests improving the technology specification's quality. Second, when outputs can be excluded yielding fewer dimensions, the estimation's quality improves leading to an increase in the

speed of convergence and a reduction in the finite sample upward bias.

To formalize this reasoning, we respecify the output vector y into two subsets of outputs, i.e. $y = \left(y^1, y^2\right)$, where $y^1 \in \mathbb{R}^{q-r}$ denotes the vector of $q-r$ outputs that are assumed to be relevant outputs of the production process under consideration, and $y^2 \in \mathbb{R}^r$ denotes the vector of r possibly redundant outputs. The hypothesis then is that x influences the level of y^1 but not of y^2. The null and alternative hypothesis can therefore be written as

$$\begin{array}{l} H_0\text{: } x \text{ influences the level of } y^1 \ \left(y^2 \text{ is redundant}\right) \\ H_1\text{: } x \text{ influences the level of } y^1 \text{and } y^2 \ \left(y^2 \text{ is relevant}\right). \end{array} \tag{5}$$

For any given input-output-combination $(x, y) = \left(x, y^1, y^2\right) \in \Psi$, the corresponding reformulated input-oriented Farrell efficiency scores in Equation 3 are:

$$\begin{array}{ll} \theta_U(x, y) & = inf\left\{\theta \mid \left(x, y^1, y^2\right) \in \Psi\right\} \\ \theta_R(x, y) & = inf\left\{\theta \mid \left(x, y^1\right) \in \Psi\right\} \end{array} \tag{6}$$

where θ_U and θ_R represent the efficiency for the unrestricted and the restricted technology set. If the outputs in y^2 are truly redundant, θ_R equals θ_U. If outputs in y^2 contain relevant outputs, then θ_R would be smaller than θ_U. From that we can derive the following inequalities:

$$\begin{array}{l} \text{if } H_0 \text{ is true: } 1 \geq \theta_U(x, y) = \theta_R(x, y), \text{ for all } (x, y) \in \Psi \\ \text{if } H_1 \text{ is true: } 1 \geq \theta_U(x, y) > \theta_R(x, y), \text{ for some } (x, y) \in \Psi \end{array} \tag{7}$$

According to Equation 4, θ_U and θ_R can be estimated from the sample, denoted by $\mathcal{X}_n$, as follows:

$$\begin{aligned} \widehat{\theta_U}(x, y) = \min_{\theta, \lambda_1, \ldots, \lambda_n} \{\theta > 0 \mid \ & \theta x \geq \textstyle\sum_{i=1}^n \lambda^i x_k^i; k = 1, \ldots, p \\ & y^1 \leq \textstyle\sum_{i=1}^n \lambda^i y_l^{1,i}; l = 1, \ldots, (q-r) \\ & y^2 \leq \textstyle\sum_{i=1}^n \lambda^i y_l^{2,i}; l = 1, \ldots, r \\ & \textstyle\sum_{i=1}^n \lambda^i = 1;\ \lambda^i \geq 0 \,\forall\, i = 1, \ldots, n\}. \end{aligned} \tag{8}$$

and

$$\begin{aligned} \widehat{\theta_R}(x, y) = \min_{\theta, \lambda_1, \ldots, \lambda_n} \{\theta > 0 \mid \ & \theta x \geq \textstyle\sum_{i=1}^n \lambda^i x_k^i; k = 1, \ldots, p; \\ & y^1 \leq \textstyle\sum_{i=1}^n \lambda^i y_l^{1,i}; l = 1, \ldots, (q-r) \\ & \textstyle\sum_{i=1}^n \lambda^i = 1;\ \lambda^i \geq 0 \,\forall\, i = 1, \ldots, n\}. \end{aligned} \tag{9}$$

where the relationship $1 \geq \widehat{\theta_U}(x,y) \geq \widehat{\theta_R}(x,y)$ holds by construction.

In order to test H_0, we have to find a valid test statistic that appropriately compares the estimated efficiencies under both technology sets. The quantity depending on the generic DGP $\mathbb{P}$ that is proposed by the literature (Simar and Wilson, 2001) is:

$$t(\mathrm{P}) = \mathrm{E}\left(\frac{\theta_U(X,Y)}{\theta_R(X,Y)} - 1\right). \tag{10}$$

From Equation 7 we know that the ratio is equal to zero, i.e. $t(\mathrm{P}) = 0$, if H_0 is true, whereas it is strictly positive otherwise, i.e. $t(\mathrm{P}) > 0$. Empirically, the ratio can easily be obtained by the sample empirical mean that is a consistent estimator (Simar and Wilson, 2001; Schubert and Simar, 2011). Therefore, the empirical equivalent of $t(\mathrm{P})$ is:

$$t_n(\mathcal{X}_n) = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{\hat{\theta}_U(X_i,Y_i)}{\hat{\theta}_R(X_i,Y_i)} - 1\right). \tag{11}$$

As mentioned before, by construction $t_n(\mathcal{X}_n) \geq 0$. Thus, the important question is how big it should be to be reasonably sure that H_0 is not true, i.e. y^2 is likely to be a relevant output of x. The usual approach is to use critical values corresponding to the distribution of the term in Equation 11. However, although this distribution can be shown to be non-degerate, it is complicated and depends on local parameters. So far the only way to determine critical values is by bootstrap-based simulation techniques. A particularly comfortable as well as flexible way is to use the subsampling approach, as suggested by Schubert and Simar (2011). This approach is described and extended to clustered data in the next subsection.

To answer the question of how large the test ratio must be in order to reject the null hypothesis, we need to compute a p-value or a critical value. This requires the approximation of the unknown (asymptotic) sampling distribution of $\tau_n(t_n(\mathcal{X}_n) - t(\mathrm{P}))$, i.e. the convergence of the test statistic $t_n(\mathcal{X}_n)$ against the true population parameter $t(\mathrm{P})$ at rate τ_n, where t_n is a function of the sample size. Note that $t_n(\mathcal{X}_n)$ is the estimate of $t(\mathrm{P})$ that discriminates between the H_0 and H_1.

The subsampling approach is a special kind of bootstrap. It differs from the normal procedure of generating pseudo samples of the original size n in that the samples here are of size $m < n$ such that $m/n \to 0$ when $n \to \infty$. This easy adjustment makes the subsampling approach robust to deviations

from the assumptions necessary for the consistency of the bootstrap. In particular, with DEA and related estimators, the frontier problem occurs, which renders bootstrapping inconsistent while the subsampling is not.

To derive an approximation of the sampling distribution of $\tau_n t_n(\mathcal{X}_n)$, we follow Schubert and Simar (2011) and use the algorithm based on subsampling proposed by Politis et al. (2001).[11] According to the algorithm, a sufficiently large number of subsets $b = 1, \ldots, B$, denoted by $\mathcal{X}^*_{m,b}$, are constructed,[12] each producing a test statistic, $t_{m,b}\left(\mathcal{X}^*_{m,b}\right)$, as defined in Equation 11. The large number of estimated test statistics approximate the sampling distribution for which a *critical value,* $\hat{t}^c_m$, can be derived. The *critical value* depends on m and the $(1-\alpha)$ quantile. At the significance level α, the test rejects H_0 if and only if the observed value is greater than the critical value, i.e. $\tau_n t_n(\mathcal{X}_n) \geq \hat{t}^c_m(1-\alpha)$, where τ_n equals $\sqrt{n}n^{2/(p+q+1)}$; for details see Schubert and Simar (2011).

We further develop the work by Schubert and Simar (2011) in the sense that we extend the applicability of the algorithm by Politis et al. (2001) to clustered data, including panel data. The panel is allowed to be unbalanced, however year-wise missing observations are assumed to be completely random. Thus, we assume away (non-random) panel selection, such as attrition. Let n be the total number of observations and n_p be the number of different companies in the panel; comparably, m and m_p are defined for the subsample case. Obviously, then $n_p \leq n$. Furthermore, for a balanced panel, L is the time length of the panel, $n_p = n/L$. In an unbalanced panel, the number of observations per company is a random integer, say Z_i, such that it has support on $0, 1, ...L$. To distinguish between the overall sample and the panel data cases, we use the subscript p whenever referring to the latter.

For company i, the test statistic in Equation 11 is then expressed as the intra-observational mean of the company-individual yearly estimates and

[11] Other bootstrap methods, e.g. the homogeneous bootstrap proposed by Simar and Wilson (1998) and further developed by Simar and Wilson (2001) or the double smooth bootstrap proposed by Kneip et al. (2008) are not applicable in our setting, because we need a method that allows for heteroscedasticity and that is valid for all data points considered simultaneously (Schubert and Simar, 2011). The aforementioned alternatives are, therefore, excluded.

[12] A large number of subsets, and hence, of subsampling replications is required in order to reconstruct the behavior of the unknown parameter. Usually, the number of replications B is set to 2,000; see e.g., Daraio and Simar (2007) and Simar and Wilson (2000).

can be rewritten as:[13]

$$t_{n_p}\left(\mathcal{X}_{n_p}, Z\right) = \frac{1}{n_p}\sum_{i=1}^{n_p}\sum_{t=1}^{L}\left(\frac{\hat{\theta}_{U,it}\left(X_i, Y_i, Z_i\right)}{\hat{\theta}_{R,it}\left(X_i, Y_i, Z_i\right)} - 1\right). \tag{12}$$

where a zero is added, if a cross-section unit is not observed in a particular year. Equation 12 differs from Equation 11 in two respects. First, the subsampling has to account for the dependence among the observations, because observations belonging to the same unit are likely to be correlated. This problem is is solved by clustering the companies across time and subsampling them block-wise as suggested by Davison and Hinkley (1997). The subsampled version of Equation 12 is then defined as

$$t_{m_{p,b}}\left(\mathcal{X}^*_{m_{p,b}}, Z^*\right) = \frac{1}{m_p}\sum_{i=1}^{m_p}\sum_{t=1}^{L}\left(\frac{\hat{\theta}_{U,it}\left(X_i, Y_i, Z_i\right)}{\hat{\theta}_{R,it}\left(X_i, Y_i, Z_i\right)} - 1\right). \tag{13}$$

Second, an additional random variable that captures the random panel response is introduced. The consistency requirement for the subsampling is that $\tau_{n_p} t_{n_p}\left(\mathcal{X}_n, Z\right)$ converges to a non-degenerated distribution (Schubert and Simar, 2011). This proof is presented in the appendix of this paper.

Irrespective of the cross-sectional or panel data case, the test procedure is sensitive to the choice of m_p, which implies a trade-off between too small and too large values. Too much information is lost if m_p is too small; if m_p is too large, the subsample size almost corresponds to the sample size n_p inducing additional biases due to inconsistency of the naive bootstrap (Daraio and Simar, 2007). Therefore, an intermediate level of m_p is supposed to balance the costs of both extremes. We use the data-driven approach, by which m_p is chosen such that the volatility of the resulting measure of interest is minimized. As volatility index we calculate the standard deviation of the 95 percent quantile of the test statistic on a running window from $m_p - 2$ to $m_p + 2$.[14] Simar and Wilson (2011) show that this data-driven approach allows for tests on m_p and on desirable power properties, e.g. rejecting H_0 with high probability when H_0 does not hold (Schubert and Simar, 2011). In order to evaluate the test statistic's volatility with respect to the choice of m_p, a grid of values m_p can reasonably take is defined. These values belong

[13] We could also normalize the inner sum by dividing by Z_i, but this will have no asymptotic effect.

[14] This corresponds to the selection rule proposed by Simar and Wilson (2008) that selects a value of m for which the resulting sample distribution and some of its features, e.g., relevant moments, are stable with respect to deviations from this particular value.

to the interval $[m_{p,min}, m_{p,max}]$. For each of these values $\hat{t}^c_{m_p}(1-\alpha)$ can be calculated and investigated with respect to their volatility. Therefore, a plot of the critical values $\hat{t}^c_{m_p}(1-\alpha)$ against the possible values of m_p reveals a first impression of where the interval's region exhibiting stable results (smallest volatilities) lies.

3.3 Outlier Detection

Since the DEA estimator envelops all observed data points to construct the frontier, it is not robust against extreme values and data errors, further referred to as outliers; see e.g., Simar (2003); Simar and Wilson (2008). Before testing the restrictions on the technology set, we perform an outlier detection procedure, using the approach suggested by Pastor et al. (1999) to identify suspicious observations. To evaluate the influence of a particular observation (say DMU_j) on the performance measure of other observations, two steps are involved: In the first step, DMU_j is removed from the sample and the efficiency estimates for all other observations are obtained as usual.

For the second step, an artificial sample is constructed that contains the observations identified as efficient in the first step plus the efficient projections of observations identified as inefficient in the first step. In the second step, the efficiency measurement program is conducted using the artificial sample and the DMU_j that was excluded in the first step. If DMU_j has no impact on the performance measurement of the remaining observations, the two efficiency estimates obtained for each of the remaining observations in the first and second step are equivalent. If both estimates differ, the remaining observations (or their efficient projections) can reduce their inputs (in the case of input-orientation) to the extent of the efficiency score obtained in the second step. We use the standard test assumptions proposed by Pastor et al. (1999), where DMU_j is considered as influential if it makes more than 5 percent of the remaining observations reduce their efficiency to less than 95 percent. Based on these parameters, a p-value can be derived, which indicates whether DMU_j should be excluded from further analysis.

4 Application to US Natural Gas Transmission Companies

4.1 Technology Specification and Variable Selection

The introduced method is applied to the sector of natural gas transmission, which is frequently subjected to regulatory benchmarking activities world-

wide. As pointed out by Jamasb et al. (2008), the regulation schemes vary among countries, with the most obvious differences between European countries and the US. Regulating natural gas transmission traditionally relies on cost-of-service or rate-of-return in the US; overviews of the implemented scheme are given, e.g., by Sickles and Streitwieser (1992, 1998) and, more recently, by O'Neill (2005). In contrast, the regulators in Europe increasingly shift toward incentive regulation, an approach discussed by e.g. Vogelsang (2002). Incentive regulation aims to introduce a company-inherent production cost reducing behavior by delegating pricing decisions to them whilst giving the opportunity to gain profits from additional cost reductions. For this purpose, incentive-based regulation typically sets price or revenue caps using the RPI-X formula (Littlechild, 1983; Beesley and Littlechild, 1989) where X is the expected saving in efficiency. The extent of the expected efficiency saving can be deduced from frontier analysis. As shown by Haney and Pollitt (2009), European regulators frequently use DEA for incentive-based regulation of the natural gas transmission companies. Although frontier analysis is currently not used to regulate US natural gas transmission companies, it is useful in this context to investigate, for instance, the total factor productivity change and technical change of the industry, particularly in the context of changing regulation; see e.g. Sickles and Streitwieser (1992); Granderson (2000); Jamasb et al. (2008).

A crucial part of both regulatory benchmarking and the evaluation of total factor productivity, *etc.*, is to specify the technology set. Consequently, extensive attention is usually devoted to the choice of variables. In their analysis on US natural gas transmission companies, Jamasb et al. (2008), for example, select the relevant variables via a comprehensive econometric cost-driver analysis. In real life applications, the conflict related to the choice of variables arises from the uncertainty about the correct specification of the technology and, in regulatory frameworks additionally, from the opposing interests of regulating authorities and regulated firms: On the one hand, firms seek to increase the number of the considered variables in order to make the model as detailed as possible and, therefore, increase the dimensions of the technology set. In the case of high dimensionality, nonparametric efficiency analysis as an regulatory instrument is compromised because no meaningful efficiency estimates can be obtained due to the 'curse of dimensionality'. Regulators, on the other hand, focus on only a few variables that appropriately model the technology set. We draw on discussions in the literature in order to establish alternative specifications of the technology set that we use to perform the proposed restriction test.

The primary task of natural gas transmission companies is to transfer

natural gas from other upstream facilities[15] to city gates, storage facilities and some large industrial customers. From the city gates on, the commodity is distributed to all other customers via local distribution systems that do not belong to the transmission system. To accomplish this task, natural gas transmission companies essentially employ pipelines, compressor stations, natural gas as fuel, and personnel.

We first specify the variables representing the inputs involved in the production process for natural gas transmission.[16] Similar to other sectors, the commonly considered input factors are i) labor; ii) "other inputs" such as e.g., fuel, materials, and power (Coelli et al., 2003); and iii) capital. The expenses on labor and "other inputs" basically constitute the operating expenses, whereas investment spending relates to capital expenses. Since compressor stations require a notable amount of fuel and maintenance, the relative share of "other inputs" is large in natural gas transmission compared to other technologies. The crucial contributors to the pipeline operating costs are, therefore, the number of compressor stations and labor expenses (IEA, 2003). With unknown factor prices, we use operating and maintenance expenses ($O\&M$) as an aggregated input measure, which sufficiently covers expenses for labor and "other inputs". The aggregated measure implies that factor prices are identical for all firms. Although this is a strong assumption, it seems reasonable in our context but must be carefully considered in each application. An advantage of the monetary aggregate is that it ensures to account for all employed inputs. In addition, from an analyst's perspective, it overcomes information asymmetries; authorities find it difficult to obtain accurate input factor prices and physical input quantities (Jamasb et al., 2008).[17]

We do not consider capital for the following reasons: First, data on capital costs or capital stock are often very limited or hardly comparable. Second, regulators frequently rely on model specifications excluding capital input related measures; see e.g., Haney and Pollitt (2009). Third, capital (or infrastructure) could alternatively be considered as a factor that enters the equation through determining the amount of labor input and "other inputs" rather than being a separate input factor. This implies for our application

[15] These mainly include gas storage facilities, gas processing and treatment plants, as well as liquefied natural gas storage and processing plants.

[16] For a general overview of commonly considered inputs and output of network industries, the reader is referred to Coelli et al. (2003); a comprehensive discussion on the variable selection in the context of gas transmission is given by e.g., Jamasb et al. (2008).

[17] Note that the legitimacy of input (or output) aggregation should also be tested, e.g. by means of restriction tests; however, this it outside the focus of the present work.

that the pipeline networks' characteristics determine how much personnel and maintenance is required to run the business.[18] Therefore, the pipeline network does not necessarily constitute an individual input.

There is a broad consensus about the plurality of outputs in network industries. The most obvious and frequently used measure to include is the natural gas delivered (*deliv*) (Coelli et al., 2003). Additionally, we consider the amount of natural gas delivered in peak times (*peak*) since the difference across firms is relevant in particular when regional characteristics vary. The provision of infrastructure (or the service supplied by using this infrastructure) itself can be considered a distinct output. Unlike other studies in which length of mains (*length*) is incorporated as some capital measure, e.g., Jamasb et al. (2008), we use it as proxy for transportation service. In addition, including *length* improves the comparability among the investigated pipeline companies. Typically, larger (existing) networks are associated with higher operational costs: Compressor stations, installed to maintain the network pressure,[19] determine a large part of personnel expenditures and maintenance costs (including fuel consumption). Not considering this technical aspect leaves companies with high *O&M* due to large networks at a disadvantage, *per se*. The network length appears to be a suitable proxy for the number of installed compressor stations since they occur at rather regular intervals of 150-200 km, corresponding to about 93-124 miles (Natgas.info, 2011).[20] Another frequently considered measure is the number of customers supplied, which accounts for the multiplicity of output. However, the number of connections seems to be of minor importance in natural gas transmission networks. We therefore exclude it from consideration. Furthermore, pollution (as a bad output) is sometimes taken into account (Coelli et al., 2003) but not considered here.

The above-mentioned discussion suggests that three potential candidates

[18] This approach however, requires additional methodological implementations that are beyond the scope of this paper.

[19] The transport of natural gas is based on a pressure differential at the inlet and outlet.

[20] However, we are aware of the fact that the length of mains cannot fully explain the differences of total operational costs of the compressor station since these also depend on the engineering characteristics. Further, length of mains likely reflects the geographical reach of services. An alternative view of its importance might result from the notion that companies active in rural areas naturally need greater *length* to deliver the same amount of gas than firms in metropolitan areas. This is simply because the customers are more dispersed. In this interpretation *length* would be rather a conditioning variable than an input or output. However, if *length* reflects an exogenous and monotonous cost disadvantage, it can also be included as an additional output. Our results are consistent with both qualifications of the variable *length* and corroborate its importance.

Table 1: Test strategy

Test	Input	Outputs	H_0	H_1	Description
I	*O&M*	*deliv*, *peak*	*deliv*	*deliv*, *peak*	*peak* is redundant
II	*O&M*	*deliv*, *peak*	*peak*	*deliv*, *peak*	*deliv* is redundant
III	*O&M*	*deliv*, *length*	*deliv*	*deliv*, *length*	*length* is redundant
IV	*O&M*	*peak*, *length*	*peak*	*peak*, *length*	*length* is redundant

for output variables to be analyzed: *deliv*, *peak*, and *length*. Given the input variable *O&M*, we develop our model in terms of a stepwise enlarging set-up as illustrated in Table 1. In Test I we test whether *peak* can be excluded from the technology set when this already includes *deliv*. In Test II we also test this, vice versa, i.e. whether *deliv* is redundant in the presence of *peak*. We find that in each of the two tests, the additional variable is redundant, but it is difficult to tell which. This finding also suggests that one measure of the delivered natural gas should be included, but it is relatively unimportant (at least empirically) which one. Given *deliv* as output variable, Test III analyzes whether *length* is an additional relevant output variable. The same is evaluated in Test IV, except for the fact that *peak* is the baseline output variable. In both latter tests we consistently reject the Null hypothesis of the possibility of exclusion.

4.2 Data

We employ data on US natural gas transmission companies provided by the FERC. FERC Form No. 2 includes all natural gas companies whose combined gas transported or stored for a fee exceed 50 mn Dth. Given we assume that the technologies of onshore and offshore pipelines differ, we consider companies operating onshore facilities only. Some missing values and data irregularities are excluded from the data set. The remaining sample contains information on 43 natural gas transmission pipeline companies that are observed with unequal frequency over a five-year time period (2003-2007).[21] In total, the unbalanced panel includes 191 observations.

By using the cross-sections over multiple years, we assume that all observations have access to the same technology, meaning that technical change is absent during the considered time span. We test this assumption using the Malmquist approach proposed by Färe et al. (1992) and investigate whether the component of technical change is neglectable. We find no empirical evi-

[21] Note that we want to empirically apply our proposed method and are, therefore, not concerned about the exact period under consideration.

dence for technical change at the 1 percent level of significance. Therefore, the observations can be considered to have access to the same technology in all years and pooling over time is valid in our case. Hence, changes in productivity are driven by productivity and technical efficiency change.

Table 2 presents the characteristics of the data. All variables are related to the companies' transmission branch. In general, all variables exhibit high standard deviations, indicating notable differences between the sample companies. Median values are consistently below corresponding mean values suggesting that the sample consists of relatively more small-size firms. For *O&M* we use the reported sum of transmission expenses for operation and maintenance. The monetary values are inflation adjusted to 2003 dollars for comparability purposes. On average the pipeline companies spend 42 mn USD on *O&M*. *Deliv* represents the account for the total quantity of natural gas delivered by the respective company and ranges from about 20 mn to 3 bn Dth. In order to ensure comparability with peak period information, we transformed this variable into Dth per day. The corresponding measure of supplied quantity then has a minimum and maximum value of 0.06 to 8.6 mn Dth a day, respectively. For *peak*, we use the single day account of the amount of natural gas delivered during system peak period. The sample companies report peak deliveries between 0.1 and 7 mn Dth (per day). *Length* represents the total length of transmission mains, which varies widely between the companies. The smallest pipeline network has 80 miles of pipeline and the largest has over 9,000 miles.

Table 2: Descriptive statistics for US natural gas transmission companies

Variable	Min	Mean	Median	Max	Std.dev.
Opex (O&M) [thsd USD[a]]	268	42,421	20,593	244,284	50,632
Total deliveries (deliv) [thsd Dth[b]]	55	1,389	994	8,597	1,381
Peak deliveries (peak) [thsd Dth]	122	1,614	1,303	7,124	1,328
Length of mains (length) [miles]	80	2,379	1,402	9,627	2,505

Source: US FERC. Notes: observations=191, n=43, years=2003-2007, onshore pipeline companies included only. [a] Yearly operating and maintenance expenses are deflated to 2003. [b] Per day measures derived by dividing the total amount of natural gas delivered by 365 days.

4.3 Results

4.3.1 Outlier Detection

First, we present the results of outlier detection. The outlier detection routine is based on a technology set that incorporates all three potential outputs simultaneously [22] and performed on a yearly base. The results are shown in Table 3 where those companies are listed that cause a loss in efficiency larger than 5 percent for at least one other company. Further, the respective number of influenced observations and the corresponding p-values are given. If the p-value is less than 10 percent, we consider the candidate to be an outlier and exclude it from further analysis. This is the case for 20 observations, thus reducing our sample for the subsequent analysis to 171 observations.

Table 3: Results of outlier detection

Year	ID	Influenced companies	p-value	Year	ID	Influenced companies	p-value
2003	22	8	0.000	2005	175	15	0.000
2003	37	22	0.000	2006	11	4	0.061
2003	78	9	0.000	2006	22	12	0.000
2003	172	11	0.000	2006	24	13	0.000
2003	175	1	0.774	2006	48	1	0.785
2004	22	25	0.000	2006	53	3	0.188
2004	53	2	0.447	2006	78	13	0.000
2004	78	12	0.000	2006	175	9	0.000
2004	175	10	0.000	2007	11	1	0.708
2005	7	1	0.774	2007	22	4	0.030
2005	22	12	0.000	2007	24	4	0.030
2005	24	14	0.000	2007	37	1	0.708
2005	43	1	0.774	2007	76	2	0.339
2005	53	3	0.175	2007	97	7	0.000
2005	78	11	0.000	2007	175	18	0.000

4.3.2 Restriction Tests

Turning to the main interest of this paper, Figure 1 illustrates the results of the restriction test for our sample based on subsampling for clustered data. The horizontal dashed lines represent the respective, actually observed values of $t_{n_p}\left(\mathcal{X}_{n_p}, Z\right)$ obtained from Equation 12. The observed test statistic is 0.3707 in Test I, 0.1907 in Test II, 2.3491 in Test III, and 2.0337 in Test IV.

[22] Calculations are conducted using the statistical software R with the additional package "FEAR" version 1.12 by Wilson (2008).

We reject the respective H_0 if this statistic exceeds the determined critical value.

To derive the empirical approximations of the sampling distributions and the corresponding critical values $\hat{t}^c_{m_p}(1-\alpha)$, we calculate 2,000 replications of the test statistic $t_{m_p,b}\left(\mathcal{X}^*_{m,b}, Z^*\right)$, for each of the four tests, using the proposed subsampling procedure. Since the critical values depend on the respective subsample sizes, the replications of the four test statistics are calculated for different values of m_p. The solid lines in the graphs illustrate the obtained corresponding critical values at the preferred level of significance (α =5 percent), as a function of the subsample size m_p.

The vertical dashed lines indicate the respective optimal subsample size determined by the smallest measured volatility index. This corresponds to a region where the test statistic graphically appears to remain stable when slightly deviating from the identified optimal value of m_p. Note that in our applied approach of subsampling for clustered data, m_p, refers to the number of cross-section covered by the subsample, not to the total number of observations. As shown by the respective panels of Figure 1, the optimal subsample sizes for our tests correspond to 39 (panel a), 36 (panel b), and 34 (panel c and d); they are the reference points where the observed values of the test statistic are compared to the critical values in order to reach a test decision.

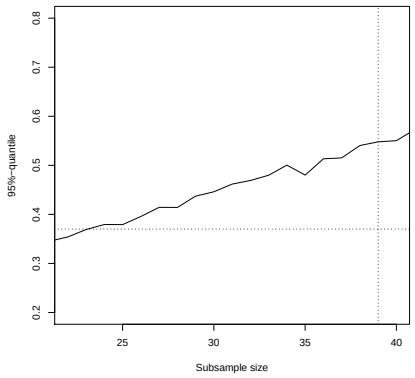
As evident from panel (a) in Figure 1, the critical value clearly exceeds the observed value of the test statistic obtained from Test I at the subsample size of 39. Therefore, we do not reject the Null hypothesis that the variable *peak* can be excluded from the technology set, given that *deliv* is an output variable. The corresponding p-value (not depicted in the graph) of this test is 46 percent, which is obviously larger than our preferred significance level of 5 percent.

Likewise there is not enough empirical evidence to reject H_0, if we run the test the other way around (Test II). At the subsample size of 36, panel (b) of Figure 1 shows that the critical value is again larger than the observed test statistic. Thus, given the output variable *peak*, the variable *deliv* is redundant to define the technology set under consideration. The p-value of Test II is slightly smaller (0.40), but roughly of the same magnitude. The results of Tests I and II indicate that we can drop either *peak* or *deliv,* if we control for the respective other one. However, the comparable significance levels show that it is relatively unimportant which is dropped. For Test III, we proceed with *deliv*, for Test IV with *peak* as the baseline output variable.

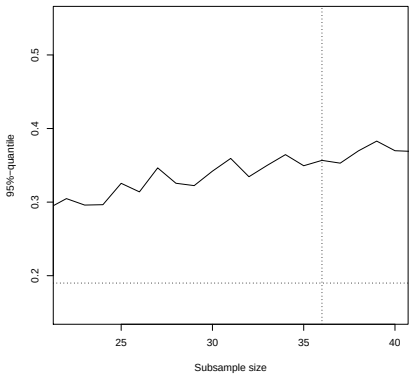
Test III compares the technology set *deliv* (H_0) against *deliv* and *length*

Figure 1: Results of restriction tests

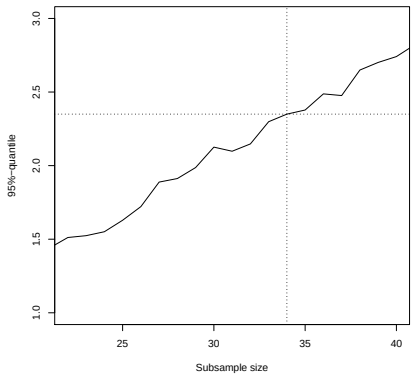
(a) Test I: Output redundancy of *peak*



(b) Test II: Output redundancy of *deliv*



(c) Test III: Output redundancy of *length* given *deliv*



(d) Test IV: Output redundancy of *length* given *peak*

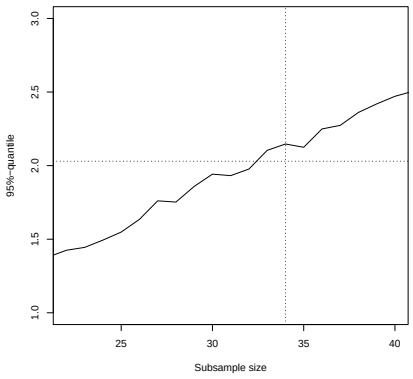


Table 4: Test statistics for Tests I to IV

Test	Optimal subsample size	Test statistic (panel)	p-value	Test statistic (pooling)	p-value
I	39	0.3707	46 %	0.4049	2%
II	36	0.1907	40 %	0.2128	11%
III	34	2.3491	4%	2.5271	0%
IV	34	2.0337	7%	2.2247	0%

(H_1). Indicated by panel (c) in Figure 1, we can indeed reject the Null hypothesis at the significance level of 5 percent level. In Test IV the relevance of *length* is marginally less pronounced at the 5 percent level. However, with a calculated p-value of 7 percent (not visible in the graph) we can reject the Null hypothesis of Test IV at the 10 percent significance level in favor of the alternative. Therefore, *length* represents a further, relevant output for the purpose of modeling the technology set of the considered companies.

Table 4 summarizes the presented results of the tests, i.e. the optimal subsample sizes, the test statistic obtained by subsampling for panel data, and the p-values. In addition, the last two columns of the table show an interesting point of comparison, i.e. the test statistics and the p-values we obtained for the presented tests, if we falsely assume that each observation is independent of each other. In this case we ignore the panel structure of the data and run the subsampling procedure without taking it into account. Technically, dependence leads to a duplication of information, because one observation already provides information about any other dependent observation. Consequently, ignoring dependence should generally overestimate the informational content of a sample leading to underestimated p-values. Indeed this is what we observe. Three out of the four tests (Tests I and III) would now reject the Null hypothesis and the only one that does not is very close to (with a p-value of 11 percent). If we based our decision on these tests, we should not exclude any of the variables, eventually sticking with the full model of *deliv*, *peak* and *length*. Note, that our results depend on the sample and do not provide general evidence for the sector. With the reduced set of outputs, we improve the efficiency estimation by reducing the dimensionality and hence, the risk of overestimating the performance.

4.3.3 Differences in the Efficiency Estimates

In practical regulation settings the estimated efficiencies have important financial implications for the companies. Each decrease in estimated efficiency

Table 5: Differences of efficiency estimates

Test	Difference between specifications	Min	Mean	Median	Max	Std.dev.
I	$\hat{\theta}_{H_1} - \hat{\theta}_{H_0}$	0.0000	0.0575	0.0154	0.4405	0.0921
II	$\hat{\theta}_{H_1} - \hat{\theta}_{H_0}$	0.0000	0.0271	0.0000	0.2744	0.0531
III	$\hat{\theta}_{H_1} - \hat{\theta}_{H_0}$	0.0000	0.2417	0.1247	0.9469	0.2766
IV	$\hat{\theta}_{H_1} - \hat{\theta}_{H_0}$	0.0000	0.2180	0.1286	0.9642	0.2493

will potentially cost the companies large amounts of money. Therefore, it is interesting to know which effects the proposed restrictions actually have on the companies's efficiency estimates.

Table 5 shows for each of the four tests how the company-individual efficiency scores respond to the difference in the technology set specification. We estimate the performance of each company using Equations 8 and 9 and calculate the differences of the achieved efficiency scores. The unrestricted models, i.e. the efficiency estimation using the technology sets under H_1, by definition provide performance measures that are equal or greater than their corresponding restricted models. For Tests I and II the technology set under H_1 is the same, i.e. involves both outputs *deliv* and *peak*. Whereas the technology set under H_0 incorporates only *deliv* in Test I and *peak* in Test II. The minimum differences of zero in both tests can be explained by the fact that there are some observations for which the efficiency score is only determined by the variable that is not excluded in the alternative technology set. Hence, excluding the other output measure does not change their performance measure. However, for some of the companies, the efficiency score changes considerably where the exclusion of output *peak* yield, for example, to a maximum difference of 44 percentage points Test I. The maximum difference in Test II is with 27 percentage points lower. The table further shows that excluding the output *peak* has, on average, greater impact than excluding the output *deliv*, i.e. the mean of the difference is 0.0575 compared to 0.0271. Given that the tests allow for the exclusion of either *deliv or peak,* from the regulator's perspective this provides a strong argument for excluding *deliv* instead of *peak*. In either way the discriminatory power is increased.

As a reference point we also present the results for Tests III and IV. However, we note that H_0 is rejected in both cases. Therefore, any potentials in higher discriminatory power are based on the fact that we falsely would restrict the technology set.

5 Conclusions

Industrial and regulatory benchmarking are commonly applied to all kinds of industries in order to improve the companies performance. Conducting such analyses requires the modeling the technology of the companies under investigation, which in practice is often a mere guess. This paper develops an approach to support the model specification of technology sets in nonparametric efficiency analysis based on statistical inference for clustered data.

To reach a decision on alternative model specifications, we propose approximating the sampling distribution of the test statistic of interest, i.e. the ratio of the efficiency estimates obtained from alternative technology sets, using a block-wise subsampling procedure. This approach ensures that the dependency between observations is properly accounted for. The corresponding critical value of the sampling distribution can subsequently be used as the decision criteria. Due to the block-wise subsampling, the applicability of restriction tests, previously only proposed for the cross-sectional case, is extended to (unbalanced and balanced) panel data structures and any other kind of dependent observations.

Panel data is, for example, particularly interesting when the relative performance is measured for a small number of units. Due to monopolistic market structures, this is the case with regulatory benchmarking of network industries. Observing the units over multiple time periods can sufficiently enlarge the sample size to obtain meaningful efficiency measures and to apply restriction tests. In addition, regulatory benchmarking involves the issue of uncertainty about the correct specification of the technology, which requires objective modeling.

Therefore, we apply and demonstrate the proposed restriction test in a regulatory framework where we consider the natural gas transmission sector. Our consecutive analysis involves four alternative technology sets for this sector, where the variable selection is based on the respective literature and regulatory practice. All technology sets in question contain operating and maintenance expenditures as input, while they differ in the output measures. The analysis is undertaken using an unbalanced panel data set of US natural gas transmission pipelines for the years 2003 - 2007.

First, we test whether the amount of natural gas delivered during peak times is a redundant output measure, if the technology set already included the total amount of natural gas delivered as an output. The second test deals with the reverse case, i.e. it tests whether the total amount of natural gas delivered is redundant, if the amount of natural gas delivered during peak times

is defined as output. The test results suggest that in each case the respective additional output variable is dispensable, meaning that the technology set is sufficiently determined by one of the output variables. Although, the test is not designed to answer the question which of the alternative output variables is the correct one to choose, further analyses on discriminatory power provide some tentative indication that peak deliverables rather than total deliverables should be included. The efficiency estimates are more sensitive toward omitting the peak amount of natural gas delivered than omitting the total amount of natural gas delivered.

Based on the first two tests, the subsequent two tests both suggest not excluding the length of mains as an output from the technology set if the initial output variable was either given by the total amount of natural gas delivered or the amount of natural gas delivered during peak times, respectively. Deleting the length of mains affects the efficiency estimates strongest indicating its importance for modeling the technology set.

For our sample, the test provides an objective tool to reduce the number of variables, which prevents overestimating the performance of the companies by including redundant variables in the specification. In general, the proposed test is a sound and reproducible method that helps removing the information asymmetry between the analyst and the production entity delivering the data and possibly being subject to regulatory benchmarking.

References

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.

Beesley, M. and Littlechild, S. (1989). The regulation of privatized monopolies in the United Kingdom. *RAND Journal of Economics*, 20(3):454–472.

Bogetoft, P. and Otto, L. (2011). *Benchmarking with DEA, SFA, and R*. International Series in Operations Research & Management Science. Springer, New York.

Coelli, T., Estache, A., Perelman, S., and Trujillo, L. (2003). A primer on efficiency measurement for utilities and transport regulators. World Bank Institute, Development Studies.

Cooper, W. W., Seiford, L. M., and Zhu, J. (2011). Data Envelopment Analysis: History, models, and interpretations. In Cooper, W. W., Seiford, L. M., and Zhu, J., editors, *Handbook on Data Envelopment Analysis*, International Series in Operations Research & Management Science, Volume 164, pages 1–39. Springer, New York, 2nd edition.

Cullmann, A. (2012). Benchmarking and firm heterogeneity: A latent class analysis for German electricity distribution companies. *Empirical Economics*, 42(1):147–169.

Daraio, C. and Simar, L. (2007). *Advanced robust and nonparametric methods in efficiency analysis: Methodology and applications*. Studies in productivity and efficiency. Springer, New York.

Davison, A. C. and Hinkley, D. V. (1997). *Bootstrap methods and their application*. Cambridge University Press, Cambridge.

Debreu, G. (1951). The coefficient of resource utilization. *Econometrica*, 19(3):273–292.

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

Färe, R. and Grosskopf, S. (2005). *New directions: efficiency and productivity*. Studies in productivity and efficiency. Kluwer Academic Publishers, Boston.

Färe, R., Grosskopf, S., Lindgren, B., and Roos, P. (1992). Productivity changes in Swedish pharmacies 1980-1989: A non-parametric Malmquist approach. *Journal of Productivity Analysis*, 3(1-2):85–101.

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

Farsi, M., Fetz, A., and Filippini, M. (2007). Benchmarking and regulation in the electricity distribution sector. CEPE Working paper series 07-54, CEPE Center for Energy Policy and Economics, ETH Zurich, Zurich.

Granderson, G. (2000). Regulation, open-access transportation, and productive efficiency. *Review of Industrial Organization*, 16(3):251–266.

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

Hollas, D. R., Macloed, K. R., and Stansell, S. R. (2002). A Data Envelopment Analysis of gas utilities' efficiency. *Journal of Economics and Finance*, 26(2):123–137.

Homburg, C. (2001). Using Data Envelopment Analysis to benchmark activities. *International Journal of Production Economics*, 73(1):51 – 58.

IEA (2003). The challenges of future cost reductions for new supply options (pipelines, LNG, GTL). 22nd World Gas Congress Tokyo, website. `http://www.dma.dk/themes/LNGinfrastructureproject/Documents/Infrastructure/IEA-The\%20challenges\%20of\%20further\%20cost\%20red\%20new\%20supply\%20options.pdf`, retrieved 26 September 2011.

Jamasb, T., Nillesen, P., and Pollitt, M. (2004). Strategic behaviour under regulatory benchmarking. *Energy Economics*, 26(5):825–843.

Jamasb, T. and Pollitt, M. G. (2003). International benchmarking and yardstick regulation: An application to European electricity distribution utilities. *Energy Policy*, 31(15):1609–1622.

Jamasb, T., Pollitt, M. G., and Triebs, T. (2008). Productivity and efficiency of US gas transmission companies: A European regulatory perspective. *Energy Policy*, 36(9):3398–3412.

Joskow, P. L. (2006). Incentive regulation in theory and practice: Electricity distribution and transmission networks. Cambridge Working Papers in Economics 0607, Cambridge University, Faculty of Economics, Cambridge.

Kneip, A., Park, B. U., and Simar, L. (1998). A note on the convergence of nonparametric DEA estimators for production efficiency scores. *Econometric Theory*, 14(6):783–793.

Kneip, A., Simar, L., and Wilson, P. W. (2008). Asymptotics and consistent bootstraps for DEA estimators in nonparametric frontier models. *Econometric Theory*, 24(06):1663–1697.

Littlechild, S. C. (1983). Regulation of British telecommunications' profitability. Report to the Secretary of State, Department of Industry in London, London.

Natgas.info (2011). Gas pipelines. Website. `http://natgas.info/html/gaspipelines.html`, retrieved 26 September 2011.

O'Neill, R. P. (2005). Natural gas pipelines. In L.Moss, D., editor, *Network access, regulation and antitrust*, pages 107–120. Routledge, London.

Pastor, J. T., Ruiz, J. L., and Sirvent, I. (1999). A statistical test for detecting influential observations in DEA. *European Journal of Operational Research*, 115(3):542–554.

Politis, D. N., Romano, J. P., and Wolf, M. (2001). On the asymptotic theory of subsampling. *Statistica Sinica*, 11(4):1105 –1124.

Schefczyk, M. (1993). Industrial benchmarking: A case study of performance analysis techniques. *International Journal of Production Economics*, 32(1):1–11.

Schubert, T. and Simar, L. (2011). Innovation and export activities in the German mechanical engineering sector: An application of testing restrictions in production analysis. *Journal of Productivity Analysis*, 36(1):55–69.

Shepard, R. W. (1970). *Theory of cost and production function*. Princeton University Press, Princeton.

Shuttleworth, G. (2005). Benchmarking of electricity networks: Practical problems with its use for regulation. *Utilities Policy*, 13(4):310–317.

Sickles, R. C. and Streitwieser, M. L. (1992). Technical inefficiency and productivity decline in the U.S. interstate natural gas pipeline industry under the National Gas Policy Act. *The Journal of Productivity Analysis*, 3(1-2):119–133.

Sickles, R. C. and Streitwieser, M. L. (1998). An analysis of technology, productivity, and regulatory distortion in the interstate natural gas transmission industry: 1977-1985. *Journal of Applied Econometrics*, 13(4):377–395.

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

Simar, L. and Wilson, P. W. (1998). Sensitivity analysis of efficiency scores: How to bootstrap in nonparametric frontier models. *Management Science*, 44(1):49–61.

Simar, L. and Wilson, P. W. (2000). Statistical inference in nonparametric frontier models: The state of the art. *Journal of Productivity Analysis*, 13(1):49–78.

Simar, L. and Wilson, P. W. (2001). Testing restrictions in nonparametric efficiency models. *Communications in Statistic: Simulation and Computation*, 30(1):159 – 184.

Simar, L. and Wilson, P. W. (2008). Statistical inference in nonparametric frontier models: Recent developments and perspectives. In Fried, H. O., Lovell, C. K., and Schmidt, S. S., editors, *The measurement of productive efficiency and productivity growth*, pages 421–521. Oxford University Press, Oxford.

Simar, L. and Wilson, P. W. (2011). Inference by the m out of n bootstrap in nonparametric frontier models. *Journal of Productivity Analysis*, 36(1):33–53.

Thompson, R. G., Dharmapala, P. S., and Thrall, R. M. (1995). Linked-cone DEA profit ratios and technical efficiency with application to Illinois coal mines. *International Journal of Production Economics*, 39(1-2):99–115.

Vogelsang, I. (2002). Incentive regulation and competition in public utility markets: A 20-year perspective. *Journal of Regulatory Economics*, 22(1):5–27.

Wilson, P. (2008). FEAR 1.0: A software package for frontier efficiency analysis with R. *Socio-Economic Planning Sciences*, 42(4):247–254.

Appendix

A robust approach to obtain corrected standard errors with clustered data is to sub-sample block-wise (Davison and Hinkley, 1997). This allows for arbitrary dependence between the observations belonging to the same cross-section unit.

We show that this procedure meets the essential consistency requirements set out in Politis et al. (2001). Let sample size n_p be defined by the number of different cross-section observations. Although we used the more easily interpretable Farrel-Debreu measure so far, for actual calculations it is preferable to use the inverse $\lambda = 1/\theta$ because it is truncated only once.

Proposition: Let $n(Z) = \sum_{i=1}^{n_p} Z_i$ where iid random variables Z_i give the number of time observations per cross-section unit with distribution function F_Z defined on the support $S_Z = 1, \ldots, L$ and expectation $c \in [1, L]$, then for the test-statistic $t_{n_p}(X, Y, Z)$ the asymptotic distribution of $\sqrt{n_p} n_p^{2/(p+q+1)} t_{n_p}(X, Y, Z)$ is non-degenerate with expectation zero.

Proof: If we reformulate the time subscripts to take only consecutive integers, we can use the following definition:

$t_i(X, Y \mid Z = z) = \sum_{t=1}^{z_i} \left(\frac{\hat{\lambda}_{U_{it}}(X_i, Y_i, \mid Z_i = z_i)}{\hat{\lambda}_{R_{it}}(X_i, Y_i, \mid Z_i = z_i)} - 1 \right)$.

It follows from the results of Kneip et al. (2008) that

$n^{2/(p+q+1)} \left(\frac{\hat{\lambda}_{U_{it}}(X_i, Y_i, \mid Z_i = z_i)}{\hat{\lambda}_{R_{it}}(X_i, Y_i, \mid Z_i = z_i)} - 1 \right) \xrightarrow{d} H_n$

for any fixed z_i, where H_n is a random variable with an asymptotic distribution function Q that is non-degenerate and has mean 0 under H_0. Furthermore we can rewrite $n = n_p c$, where c is the expectation of Z. Replacing and rearranging yields

$n_p^{2/(p+q+1)} \left(\frac{\hat{\lambda}_{U_{it}}(X_i, Y_i, \mid Z_i = z_i)}{\hat{\lambda}_{R_{it}}(X_i, Y_i, \mid Z_i = z_i)} - 1 \right) \xrightarrow{d} \frac{1}{c^{2/(p+q+1)}} H_n.$

Since the right-hand-side is a scaled version of H_n, also

$n_p^{2/(p+q+1)} \left(\frac{\hat{\lambda}_{U_{it}}(X_i, Y_i, \mid Z_i = z)}{\hat{\lambda}_{R_{it}}(X_i, Y_i, \mid Z_i = z)} - 1 \right)$

has a non-degenerate distribution. This implies that the conditional distribution of $n_p^{2/(p+q+1)} t_i(X, Y | Z = z)$ is non-degenerate. Call this distribution $D(z)$.

Furthermore, we obtain the distribution of $t_i(X, Y, Z)$ by marginalizing out Z: $D(\cdot) = \int_{z \in S_Z} D(z) dF_Z$. Obviously, if $D(z)$ is non-degenerate with a

given scaling factor, then $D(\cdot)$ must be non-degenerate with the same scaling factor. In order to complete the proof, since $t_{n_p}(X, Y, Z)$ is an empirical mean of the $t_i(X, Y, Z)$, it follows by using the redefinition $n = n_p c$ that $\tau_{n_p} t_{n_p(X,Y,Z)}$ with $\tau_{n_p} = \sqrt{n_p} n_p^{2/(p+q+1)}$ is non-degenerate and additionally has an asymptotic expectation equal to zero under H_0, because the mean associated with the asymptotic distribution Q is zero. As a consequence of this result, the subsampling methods proposed by Politis et al. (2001) are consistent, when subsampling is conducted block-wise along the cross-section dimension. The sub-sampling size m_p is as usually defined as the integer part of n_p^k for $0 < k < 1$. It should be noted that these results include the case of ordinary cross-section data and a balanced panel setting. In the former case $z_i = 1$ and $n = n_p$ yielding just the formulae in Schubert and Simar (2011). In the latter case $z_i = L$ implying that z_i cannot affect the asymptotic distribution because it is non-random.