Testing Scale Efficiency in DEA Models: A Smooth Bootstrap Approach

Hédi Essid
Institut Supérieur de Gestion
41, Rue de la liberté, Cité Bouchoucha 2000
Le Bardo, Tunis-Tunisie

Pierre Ouellette
Department of Economics,
Université du Québec à Montréal
PO Box 8888, Station Centre-Ville
Montréal, Québec

Stéphane Vigeant
IESEG School of management
&
EQUIPPE (Universités de Lille)
Faculté des Sciences Économiques et Sociales
Université des Sciences et Technologies de Lille
59655 Villeneuve d'Ascq Cédex, France

Abstract

This paper discusses and implements a nonparametric statistical test procedure for organization scale efficiency. This procedure allows the practitioner to test whether the observed scale efficiency score is real or due to sampling variation. Our test procedure is based on the work of Simar and Wilson (1998, 2002). However, instead of considering a global test to give an average measure of the efficiency of a sector of activity, we consider an individual test which gives a measure of the scale efficiency for each organization. We believe that this type of analysis is the most used in empirical studies. The test is simulated in a Monte Carlo study and is shown to have the correct size in most cases. In contrast, the shape of the estimated power function is not expected. This can be explained by the "curse of dimensionality" problem which is often raised in this context.

Keywords: Returns to scale, Data envelopment analysis, Bootstrap, Monte-Carlo, Test of Hypothesis.

JEL Classifications: C13; C14; D20

1. Introduction

In this paper, we address the question of the qualitative measurement of returns to scale in organizations. This measure allows us to determine if organizations' activities are characterized by either increasing, decreasing or constant returns to scale. Our approach is anchored in the DEA tradition. Returns to scale have received substantial attention, leading to three measurement methods under the DEA approach: the scale efficiency method (see Färe and Grosskopf (1985) and Färe *et al.* (1994), the sum of intensity variables method of Banker (1984), Banker and Thrall (1992) and Färe and Grosskopf (1994), and the dual variable-sign-method of Banker *et al.* (1984). Banker *et al.* (1994) have shown that these three methods are equivalent. It is not uncommon however, that practitioners using these methods neglect the statistical content of the efficiency measures and do not distinguish between the true value and the estimator.

If the data are generated by a distribution with bounded support, based on the real production set, then the scale efficiency is measured with respect to an estimated frontier with data obtained from an unobserved Data Generating Process (DGP). Therefore, the observed organization's efficiency is noisy and spoiled by statistical noises. That is, efficiency measurements are subject to sampling variations. Consequently, it is necessary to adopt a formal statistical methodology that differentiates between the real efficiency value and its estimator in order to evaluate the statistical significance of results concerning the returns to scale of an organization.

Banker (1993, 1996) proposed tests of model structure based on ad-hoc distributional assumptions. However, Kittelsen (1997) raises two more problems with this approach. First, the efficiency measure estimators are biased in small sample (finite sample) and converge at a very slow rate when the dimensionality of the problem is large. Second, the components of the scale efficiency score estimators are correlated and consequently can never follow asymptotically the Fisher distribution found by Banker. Löthgren and Tambour (1999) used the approach developed by Färe *et al.* (1994) to measure the scale efficiency and address the implementation of a global test of the returns to scale. They develop a test procedure based on a *naïve* bootstrap procedure to approximate the test statistic distribution under the null hypothesis. However, Simar and Wilson (1999, 2000a) have shown that the *naïve* bootstrap method does not converge for non parametric estimators of efficiency measures of the DEA type. Simar and Wilson (2002) develop a test procedure based on the smooth bootstrap method. This method consists in smoothing the probability distribution of the efficiency scores with a convolution kernel. Then the authors use a reflection method (Silverman, 1986 and Schuster, 1985) to correct the estimator of the probability density in the neighborhood of the boundaries.

Recently, Kneip *et al.* (2008, 2009) developed a consistent bootstrap method for inference about the efficiency of a single, fixed point. However, this method requires complicated coding and is not appropriate for approximating the sampling distribution of a test statistic (Simar and Wilson, 2011). In this article, we adopt the approach developed by Simar and Wilson (2002) to handle the statistical test procedure, but we consider a specific test problem for each organization instead of a global procedure. This means that our test procedure provides an estimate of the scale efficiency for each organization. The problem of test is repeated in this case, as many times as the number of organizations.

The performance of the test procedure is assessed through a Monte Carlo experiment. We study, at first the reliability of the test. Our simulation experiments will focus on calculating the difference between the nominal and the real size in case of a technology of one input and one output, therefore in case of two inputs and one output. However, the reliability condition is not sufficient. The hypothesis tested may be true, but it may as well be false. For this reason we conduct simulation experiments in order to estimate the power curve for a technology of one input and one output.

2. Production technology and returns to scale

Consider a production activity that use input quantities $x \in \mathbb{R}_+^m$ to produce the output quantities $y \in \mathbb{R}_+^s$. It is common practice in economic analysis to describe the activity of a productive activity by means of the production set Ψ of physically attainable points (x, y):

$$\Psi = \left\{ (x, y) \in \mathbb{R}_+^{m+s} \middle| (x, y) \text{ is feasible} \right\}.^{1} \quad (1)$$

To characterize the returns to scale environment of the firms and to test for their type, specific production sets are required. The returns to scale are characterized by the way one can either expand the scale in the production set or shrink it or both. That is, the technology exhibits non increasing (*nirs*), non decreasing (*ndrs*) or constant returns to scale (*crs*) depending on the values assumed by the positive scalar α in the following set:

$$\Psi^k = \left\{ (x, y) \in \Psi \middle| (\alpha x, \alpha y) \in \Psi \text{ for all } \alpha \in \mathrm{K}^k \right\}, \; k = \textit{nirs, ndrs} \text{ and } \textit{crs} \quad (2)$$

where $\mathrm{K}^{nirs} = [0,1)$, $\mathrm{K}^{ndrs} = [1,\infty)$ and $\mathrm{K}^{crs} = [0,\infty)$. A technology that exhibits *ndrs*, *nirs* or *crs* in different regions of the production frontier is said to be characterized by variable returns to scale (*vrs*). We denote this production possibility set by Ψ^{vrs}.

It is possible to define an input oriented technical efficiency measure in the sense of Farrell (1957) with respect to the various assumptions concerning the returns to scale. That is:

$$\theta^k (x, y) = \min \left\{ \theta \middle| (\theta x, y) \in \Psi^k \right\}, \; k = \textit{nirs, ndrs, crs} \text{ and } \textit{vrs} \quad (3)$$

[1] We suppose that the production set Ψ is closed, satisfies free disposal of inputs, is bounded for finite inputs and positive outputs require positive inputs. These are standard assumptions on the technology and are discussed in Shephard (1970) and Färe (1988), among others.

From Färe and Grosskopf (1985), the (technical) efficiency measures defined in (3) can be used to construct scale efficiency measures for each organization. A scale efficiency measure is the ratio of the efficiency measure under *crs* technology and a *vrs* type technology. That is:

$$S_1(x,y)=\frac{\theta^{crs}(x,y)}{\theta^{vrs}(x,y)}\leq 1 \qquad (4)$$

We say that the production technology is of the *crs* type if $S_1(x,y)=1$. To determine the returns to scale of the technology when $S_1(x,y)<1$ we compute a second ratio that is less restrictive than the initial ratio. That is:

$$S_2(x,y)=\frac{\theta^{nirs}(x,y)}{\theta^{vrs}(x,y)}\leq 1 \qquad (5)$$

If $S_2(x,y)=1$ we say that the technology is characterized by decreasing returns to scale. Otherwise, i.e. $S_2(x,y)<1$, we say that the technology exhibits increasing returns to scale.

3. Test of hypothesis for returns to scale

In empirical research, the true model of Section 2 is unknown, so one can only aspire to approach it as closely as possible. Consequently, to identify and to characterize the returns to scale exhibited by a technology, we have to construct a statistical model to estimate the necessary parameters and then construct a formal hypothesis test procedure. We can derive the test procedure from Section 2 as a sequence of two sub-tests of hypothesis. These are:

$$Test\#1: H_0: S_1(x,y)=1$$
$$H_A: S_1(x,y)<1$$

If the test concludes that we reject the null hypothesis, then we conduct the following test:

$$Test\#2: H_0^{'}: S_2(x,y)=1$$
$$H_A^{'}: S_2(x,y)<1$$

To implement this procedure we need to estimate the technology with a consistent statistical model and then deduce the test and construct the appropriate decision criteria for hypothesis testing. This is done in the remaining of this section.

3.1 Efficiency Scores Estimation

Since Ψ is unknown, it must be estimated from a sample $\Psi_0=\left\{\left(x_j,y_j\right)\right\}_{j=1}^{n}$ of data on firms' input and output quantities. The smallest convex envelop of the data gives the DEA estimator in the *vrs* case:

$$\hat{\Psi}^{vrs}=\left\{(x,y)\in\mathbb{R}^{m+s}\,\middle|\,x\geq\sum_{j=1}^{j=n}\lambda_j x_j, y\leq\sum_{j=1}^{j=n}\lambda_j y_j, \sum_{j=1}^{j=n}\lambda_j=1\ \forall\ j=1,\ldots,n\right\}. \qquad (6)$$

To obtain the estimators for the other types of returns to scale, $\hat{\Psi}^{nirs}$, $\hat{\Psi}^{ndrs}$ and $\hat{\Psi}^{crs}$ it is sufficient to slightly alter the constraint on the sum of the λ_j. That is, in the *nirs* case we have $\sum_{j=1}^{j=n}\lambda_j\leq 1$, and in the *ndrs* $\sum_{j=1}^{j=n}\lambda_j\geq 1$ must hold and finally, to obtain a *crs* envelop, no constraints on the λ_j are necessary other than non negativity of the λs.

Substituting $\hat{\Psi}^k$ for Ψ^k in (3) yields the DEA estimator of $\theta^k(x,y)$ given by:

$$\hat{\theta}^k(x,y)=\min\left\{\theta\,\middle|\,(\theta x,y)\in\hat{\Psi}^k\right\},\ k=\textit{nirs, ndrs, crs}\text{ and }\textit{vrs} \qquad (7)$$

The estimators of the scale efficiency score are naturally defined as $\hat{S}_1(x,y) = \hat{\theta}^{crs}(x,y) / \hat{\theta}^{vrs}(x,y)$ and $\hat{S}_2(x,y) = \hat{\theta}^{nirs}(x,y) / \theta^{vrs}(x,y)$.[2]

To insure the consistency of the estimator, it is necessary to specify a statistical model that allows us for a full to characterization of the data generating process (DGP). This is the content of the following assumption.

Assumption 1: The set of observations $\{(x_j, y_j)\}_{j=1}^{n}$ are identically and independently distributed (*i.i.d.*) random variables with probability density function $f(x,y)$ defined on the convex attainable set Ψ.

The construction of the statistical model is easier when we use the polar coordinate for the inputs instead of the usual Cartesian representation. To each input vector *x* we can associate a unique angle and radius. Thus, the density has two equivalent representations, $f: (x,y) \Leftrightarrow (\theta, \eta, y)$. This allows us to decompose the joint density as follows:

$$f(\theta,\eta,y) = f_\theta(\theta|\eta,y)\, f_\eta(\eta|y)\, f_y(y), \tag{8}$$

where all densities are well defined.

Assumption 2: The probability density function $f(x,y)$ is continuous on the interior of Ψ and $f(x^\partial(z,y), z, y) > 0$ where $x^\partial(z,y) = \theta(x,z,y)x$ is any point on the frontier of Ψ.

Assumption 3: The efficiency measure $\theta(x,y)$ is differentiable in *x* and *y*.[3]

Assumption A1, A2 and A3 together define the statistical model that allows us to characterize the DGP, denoted $\Im$. In fact, the DGP is entirely characterized by the production possibility set Ψ and the density *f*. That is,

$$\Im = (\Psi, f(x,y)) = (\Psi, f(\theta,\eta,y)). \tag{8}$$

This statistical model ensures the convergence of the estimator and permits a characterization of the asymptotic properties of nonparametric estimators of the DEA type. The consistency of the estimator $\hat{\Im} = (\hat{\Psi}, \hat{f})$ is studied in Korostelev *et al.* (1995a, 1995b), Banker (1993), Kneip *et al.* (1998), among others.[4]

Although DEA estimators have been widely used, inference about the underlying model structure or the efficiencies that are estimated remains problematic. Gijbels *et al.* (1999) derived the asymptotic distribution of a DEA efficiency estimators in the case of one input and one output. Jeong (2004) derived the limiting distribution of DEA efficiency estimators under *vrs* for the special case $m = 1, s \geq 1$ in the input orientation (or $m \geq 1, s = 1$ in the output orientation). Kneip *et al.* (2008) and Park *et al.* (2010) derived the limiting distributions of DEA efficiency estimators under *vrs* and *crs* (respectively) in a multivariate framework. However, these distributions contain several unknown quantities, and are not useful in a practical sense for inference (Simar and Wilson, 2011). In this situation, the bootstrap method seams to offer a solution to make statistical inference on the basis of the DEA estimator. Simar and Wilson (1998, 2000b) proposed bootstrap methods for inference about efficiency based on DEA estimators in a multivariate framework, but consistency of these procedures has not been established.

---

[2] The DEA model adopted in this paper is input oriented. Obviously, the projection point may impact on the exact measurement of the RTS. Consequently, an output oriented model may lead to a different assessment of the returns to scale than the one obtained under an input oriented model. However important this problem is, the procedure developed here cannot discriminate between these characterizations as the statistical models these DEA estimators are not nested. The procedure proposed here can be translated to the output oriented case, however.

[3] This assumption gives the necessary smoothness of the frontier function. Kneip *et al.* (1998) have shown that the convergence rate of the estimator depends on this smoothing condition. However, the smoothing condition we have used is slightly stronger than the one required for convergence. As shown by Kneip *et al.* (1998), it is sufficient that the efficiency measure satisfy a Lipschitz condition.

[4] See Simar and Wilson (2000a) for a survey of the statistical properties of DEA estimators.

Moreover, Simar and Wilson (1998) incorporated a strong assumption about the homogeneity of the inefficiency process. Kneip *et al.* (2008) proposed two bootstrap procedures for inference about efficiency, and proved consistency of both methods. However, the method requires solving *n* auxiliary linear programs for each of *B* bootstrap replications, leading to a formidable computational burden. Recently, Kneip *et al.* (2009) developed a consistent bootstrap method for inference about the efficiency of a single, fixed point while avoiding the computational burden of Kneip *et al.* (2008). However, Simar and Wilson (2011) show that the method of Kneip *et al.* (2009) is not appropriate for approximating the sampling distribution of a test statistic to different points in the sample space. In this paper, we adopt the approach of Simar and Wilson (1998). Despite the refinements made to the bootstrap method, the smooth bootstrap of Simar and Wilson is easy to implement, to generate pseudo-samples and approximate the sampling distributions.

3.2 Bootstrapping Test Statistics for Returns to Scale

With the statistical model above, we are now in a position to give a formal procedure to test returns to scale. The problem at hand to test the null hypothesis $H_0 : \omega = \omega_0$ against an alternative hypothesis of the following form: $H_A : \omega < \omega_0$. In our problem, the parameter ω is represented by the ratio $S_1(x,y)$ for the first test and $S_2(x,y)$ for the second. For both test procedures, the null hypothesis is $H_0 : \omega_0 = 1$ and for both tests we also have a consistent estimator $\hat{\omega}$ of ω: $\hat{\omega} = \hat{S}_1(x,y)$ for $\omega = S_1(x,y)$ and $\hat{\omega} = \hat{S}_2(x,y)$ for $\omega = S_2(x,y)$. In other words, the procedure consists in inferring whether or not $\omega = 1$ by measuring the distance between the estimate of ω and 1 and using a criterion to determine if the distance is small enough to conclude that they are in fact equal. To do this we need the distribution of the test statistic $(\hat{\omega} - \omega_0)$ under the null hypothesis, H_0. This statistical distribution is difficult to define. We use in this case, the bootstrap methodology. This method consists in replicating identically the DGP many times, let's say *B* times. This allows us to generate, from the original sample Ψ_0, *B* pseudo-samples: $\Psi_b^*, \quad b = 1, \ldots, B$. Then, the original estimation method (DEA in our case) is applied to each pseudo-samples to obtain the bootstrap estimator of the test statistic $\hat{\omega}_b^*$ of $\hat{\omega}$. This procedure allows us to estimate the empirical distribution of $(\hat{\omega}^* - \hat{\omega})$ used to approximate the unknown distribution of the statistic $(\hat{\omega} - \omega_0)$ under H_0. That is,

$$(\hat{\omega} - \omega) \big| H_0, \Im \overset{approx}{\square} (\hat{\omega}^* - \hat{\omega}) \big| H_0, \hat{\Im} \tag{9}$$

where $\hat{\Im}$ is a consistent estimator of the DGP $\Im$ under the hull hypothesis $H_0 : \omega = 1$.

To obtain the pseudo-scores $\hat{\omega}^*$ we have to identify the correct bootstrap estimators $\hat{\theta}^{*crs}$, $\hat{\theta}^{*vrs}$ and $\hat{\theta}^{*nirs}$ of $\hat{\theta}^{crs}$, $\hat{\theta}^{vrs}$ and $\hat{\theta}^{nirs}$ respectively. To obtain their estimators, we use a smooth bootstrap methodology developed by Simar and Wilson (1998) and we suppose that each pseudo-efficiency is generated from a DGP under the same type of returns to scale than the initial efficiency score. This implies that the returns to scale should be constant for the first test and the returns to scale should be non-increasing for the second test. This ensures that the pseudo-scores take permissible values (less than 1). Therefore, equation (9) under H_0 must be rewritten as follows:

$$(\hat{S}_1 - 1) \big| \Im^{crs} \overset{approx}{\square} (\hat{S}_1^* - \hat{S}_1) \big| \hat{\Im}^{crs} \text{ for Test \# 1} \tag{10}$$

and

$$(\hat{\eta} - 1) \big| \Im^{nirs} \overset{approx}{\square} (\hat{\eta}^* - \hat{\eta}) \big| \hat{\Im}^{nirs} \text{ for Test \# 2.} \tag{11}$$

Our approach is based on Simar and Wilson (2002) as we use the same smooth bootstrap methodology to generate the pseudo-scores. However, we consider a specific test problem for each organization instead of a global test. This means that the problem of test is repeated as many times as the number of DMU, *n*.

To generate the pseudo-efficiencies $\hat{\theta}^{*crs}$, $\hat{\theta}^{*vrs}$ and $\hat{\theta}^{*nirs}$ we use a homogenous bootstrap methodology developed by Simar and Wilson (1998). That is, the efficiency score θ is independent of (η, y): $f(\theta|\eta, y) = f(\theta)$.

A consistent estimator of *f*, obtained using a kernel estimator and corrected by Schuster's (1985) and Silverman (1985) is defined as follows:

$$\hat{f}^{c}(t) = \begin{cases} 2\hat{g}(t) & \text{if } t \le 1 \\ 0 & \text{otherwise} \end{cases}, \text{ where } \hat{g}(t) = \frac{1}{2nh}\sum_{j=1}^{j=n}\left[\phi\left(\frac{t-\hat{\theta}_j}{h}\right) + \phi\left(\frac{t-2+\hat{\theta}_j}{h}\right)\right] \qquad (12)$$

We use a normal Gaussian kernel, denoted ϕ, and the bandwidth, *h*, is set following the normal reference rule (Silverman (1986)). To generate the pseudo-scores $\hat{\theta}^{*k}, k = crs, vrs, nirs$ from $\hat{f}^{c}$, we propose an algorithm composed of five steps:

Step 1: Compute $\hat{\theta}_j^{crs} = \hat{\theta}_j^{crs}(x_j, y_j)\ \forall\, j = 1, \ldots, n$ using equation (7).

Step 2: Generate smoothed resampled pseudo-efficiencies as follows. First generate $\{\rho_j^*, j = 1, \ldots, n\}$ by resampling with replacement a sample of size *n*, from the empirical distribution $\{\hat{\theta}_j^k, j = 1, \ldots, n\}$. Then generate the sequence $\{\tilde{\rho}_j^*, j = 1, \ldots, n\}$ as follows:

$$\tilde{\rho}_j^* = \begin{cases} \rho_j^* + h\varepsilon_j^* & \text{if } (\rho_j^* + h\varepsilon_j^*) \le 1 \\ 2 - (\rho_j^* + h\varepsilon_j^*) & \text{otherwise} \end{cases}, \text{ where } \varepsilon_j^* \sim N(0,1).$$

Then, generate the pseudo-efficiencies γ_j^* using $\gamma_j^* = \bar{\rho}^* + (\tilde{\rho}_j^* - \bar{\rho}^*)/\sqrt{1 + h^2/\hat{\sigma}_{\hat{\theta}^k}^2}$ where $\bar{\rho}^* = (1/n)\sum_{j=1}^{n}\rho_j^*$.

Step 3: Compute the pseudo variable inputs, $x_j^* = (1/\gamma_j^*)\hat{\theta}_j^k x_j$.

Step 4: Compute the bootstrapped efficiency measures $\hat{\theta}_j^{*crs}$ using the pseudo variable inputs based on the following program:

$$\hat{\theta}^{*crs}(x, y) = \min\left\{\theta | \theta x \ge \sum_{j=1}^{j=n}\lambda_j x_j^*,\ y \le \sum_{j=1}^{j=n}\lambda_j y_j, \lambda_j \ge 0\right\};$$

and

$$\hat{\theta}^{*vrs}(x, y) = \min\left\{\theta | \theta x \ge \sum_{j=1}^{j=n}\lambda_j x_j^*,\ y \le \sum_{j=1}^{j=n}\lambda_j y_j, \sum_{j=1}^{j=n}\lambda_j = 1,\ \lambda_j \ge 0\right\}.$$

Step 5: Repeat steps 2-5 *B* times to obtain *B* efficiency measures $\{\hat{\theta}_{jb}^{*k}, j = 1, \ldots, n, b = 1, \ldots, B\ k = vrs, crs\}$.

If the test concludes that we reject the null hypothesis, then one conduct the second test. To generate the data in this second case, we use the following:

Step 1: Compute $\hat{\theta}_j^{nirs} = \hat{\theta}_j^{nirs}(x_j, y_j)\ \forall\, j = 1, \ldots, n$ using equation (7).

Steps 2 and 3: As above.

Step 4: Compute the bootstrapped efficiency measures $\hat{\theta}_j^{*nirs}$ and $\hat{\theta}_j^{*vrs}$ using the pseudo variable inputs based on the following program:

$$\hat{\theta}^{*nirs}(x, y) = \min\left\{\theta | \theta x \ge \sum_{j=1}^{j=n}\lambda_j x_j^*,\ y \le \sum_{j=1}^{j=n}\lambda_j y_j, \sum_{j=1}^{j=n}\lambda_j \le 1,\ \lambda_j \ge 0\right\};$$

And

$$\hat{\theta}^{*vrs}(x,y) = \min\left\{\theta | \theta x \geq \sum_{j=1}^{j=n} \lambda_j x_j^*,\ y \leq \sum_{j=1}^{j=n} \lambda_j y_j, \sum_{j=1}^{j=n} \lambda_j = 1,\ \lambda_j \geq 0\right\}.$$

Step 5: Repeat steps 2-5 *B* times to obtain *B* efficiency measures $\left\{\hat{\theta}_{jb}^{*k}, j = 1,\ldots,n, b = 1,\ldots,B, k = vrs, nirs\right\}$.

We replicate *B* times the statistic $\hat{S}_{1j} = \hat{\theta}_j^{crs} / \hat{\theta}_j^{vrs}$ for each (x_j, y_j) in the sample set Ψ_0. Then, we obtain *B* pseudo-values $\hat{S}_{1jb}^* = \hat{\theta}_{jb}^{*crs} / \hat{\theta}_{jb}^{*vrs}, b = 1,\ldots,B$ for all $j = 1,\ldots,n$. For the second test, the application of the bootstrap procedure on the statistic $\hat{S}_{2j} = \hat{\theta}_j^{nirs} / \hat{\theta}_j^{vrs}$ leads to *B* replications: $\hat{S}_{2jb}^* = \hat{\theta}_{jb}^{*nirs} / \hat{\theta}_{jb}^{*vrs}, b = 1,\ldots,B$ for each production plan in Ψ_0.

Once the pseudo-scores $\hat{\omega}^*$ are generated, we proceed to the decision rule for each organization. The true critical value of the test satisfies $\Pr\left(\hat{\omega} - \omega_0 \leq -c_\alpha \middle| H_0 : \omega = \omega_0\right) = \alpha$. Since we do not observe this distribution, we use the bootstrap analog, $\Pr\left(\hat{\omega}^* \leq \hat{\omega} - c_\alpha^* \middle| H_0, \hat{\Im}\right) = \alpha$ where the equality holds asymptotically. In practice, c_α^* is an approximation of the true critical value, c_α. To obtain c_α^*, we start by sorting in ascending order the pseudo-scores $\left(\hat{\omega}_b^* - \hat{\omega}\right)$ for $b = 1,\ldots,B$. Then, we eliminate $(1-\alpha)\times 100$ percent of the values at the right end of the ordered sequence of bootstrapped statistics. The critical value $-c_\alpha^*$ is then the right end of this truncated sequence. Because we use a limited bootstrap experiment, we can only obtain the following approximation by substituting this bootstrapped c_α^*:

$$\Pr\left(\hat{\omega} \leq \omega_0 - c_\alpha^* | H_0\right) \approx \alpha \qquad (13)$$

Then, the decision rule is the following: reject the null hypothesis $H_0 : \omega_0 = 1$ if, for a given size α, $\hat{\omega} \leq 1 - c_\alpha^*$. Hypothesis tests can be more convincingly used by giving the marginal probability of rejections, or *p*-value. Let $\hat{\omega}_{obs}$ denote a realized value of the test statistic $\hat{\omega}$. The *p*-value for H_0 is defined to be $p = \Pr\left(\hat{\omega} \leq \hat{\omega}_{obs} \middle| H_0 : \omega = \omega_0\right)$. Under H_0, this is equivalent to $p = \Pr\left(\hat{\omega} - \omega_0 \leq \hat{\omega}_{obs} - \omega_0 \middle| H_0\right)$. The bootstrap analogue of this last expression is:

$$\hat{p} = \Pr\left(\hat{\omega}^* - \hat{\omega} \leq \hat{\omega}_{obs} - \omega_0 \middle| H_0, \hat{\Im}\right) \qquad (14)$$

Note that we have $\hat{\omega} = \hat{\omega}_{obs}$ and $\hat{\omega}$ is a consistent estimator of ω_0. Then, equation (14) is asymptotically equivalent to the following probability:

$$\hat{p} = \Pr\left(\hat{\omega}^* \leq \hat{\omega}_{obs} \middle| H_0, \hat{\Im}\right) \qquad (15)$$

If α is the size of the test, then we reject H_0 when $\hat{p} \leq \alpha$.

4. A Monte Carlo experiments

In this section we conduct a series of Monte Carlo experiments to evaluate the performance of our test procedure.[5] The experiments are conducted under the following general framework. At each Monte Carlo iteration we use two thousand Bootstrap replications, *B*=2000, for each units and the experiment is repeated one thousand times, *N*=1000. In each Monte Carlo experiment, we compute the test statistic for each DMU, using the data generated for the inputs and the outputs. Then we use the bootstrap procedure presented in the previous section to determine the p-value for each DMU. For each Monte Carlo experiment, we estimate the real size of each test as the number of times the null hypothesis is rejected divided by the number of experiments (*N*), given a nominal size (the theoretical □). For the bootstrap simulations in each Monte Carlo experiment, we use a normal density kernel and Silverman normal rule to calculate the bandwidth.

[5] The Monte Carlo experiment uses a SAS program written by the authors.

The diagnostic of the performance of the tests with respect to the size is easy to understand: If the theoretical size is correct, the real size calculated from the Monte Carlo experiment must be in the neighborhood of this true value. We conduct a total of eight experiments. They are all with one output (*s*=1), but we consider either one or two inputs, $\left(m \in \{1,2\}\right)$, samples of DMUs of size ten and twenty, $n \in \{10,20\}$, and all experiments are ran for test sizes $\alpha = 0,05$ and $\alpha = 0,01$.

We consider the performance of both tests separately. Let us start with an assessment of the performance of the first test. In order to have a tractable problem, easy to understand, we have used a simple data generating process where the number of output is set to one in all experiments and the number of inputs is set to either one or two depending on the experiment. The inputs of DMU *j*, $\left(x_{ij}\right)$, are assumed to be independently and identically uniformly distributed (*iid*) on the interval $[1,9]$. To generate the output under the null hypothesis of constant returns to scale, we proceed as follows. First we generate a sequence of *n* independently distributed standard normal noises, $N(0,1)$, v_j for *j*=1,...,*n*. Then, this noise is used to generate the output for the j^{th} DMU, $\left(y_j\right)$, according to the following constant returns to scale Cobb-Douglas production structure:

$$y_j = \prod_{i=1}^{m} x_{ij}^{1/m} e^{-0.1|v_j|},\ j = 1,\ldots,n \qquad (16)$$

Table 1 summarizes the results of the eight Monte-Carlo experiments related to the performance of test #1. A comparison of the real size from the Monte-Carlo experiment and the nominal size (the theoretical size, α), shows that the results are close to the expected size at both one and five percent and for both sample sizes, confirming the validity of our test procedure. The first part of the table presents the results for *n*=10. On average, the real size is equal to 0,0443 for $\left(m = s = 1\right)$, a number fairly close to the nominal size of $\alpha = 0,05$. The distortion between the sizes is null for DMU 5 and negligible for DMU 1 and 6 with a real size of 0,051 and 0,049 respectively. When we increase the number of inputs to two, $\left(m = 2\right)$, the average real size is now equal to 0,0395. That is, the distortion in the size increases, but this can probably be explained by the curse of dimensionality. Similar remarks apply for the size $\alpha = 0,01$. The real size is equal to 0,0142 for the one-input experiment (*m=1*) and equal to 0.0128 for two-input experiment (*m=2*), which are still very close to the nominal size.

When we increase the sample size to twenty, $n = 20$, the distortions decrease significantly, to almost 0,001 for both nominal sizes and both experiments with the number of inputs set to either one $\left(m = 1\right)$ or two $\left(m = 2\right)$. This clearly shows that our test procedure is performing better when the sample size increases, as it would be expected.

Table 1: Monte-Carlo estimates of size
Test #1 (H_0 : CRS)

DMU	Nominal size			
	$\alpha = 0,05$		$\alpha = 0,01$	
	$m = s = 1$	$m = 2,\ s = 1$	$m = s = 1$	$m = 2,\ s = 1$
n=10				
1	0,051	0,054	0,021	0,020
2	0,035	0,029	0,012	0,003
3	0,038	0,033	0,011	0,017
4	0,039	0,034	0,018	0,012
5	0,050	0,043	0,017	0,014
6	0,049	0,051	0,015	0,017
7	0,042	0,023	0,012	0,005
8	0,040	0,049	0,007	0,015
9	0,057	0,037	0,013	0,012
10	0,042	0,042	0,016	0,013
Mean	0,0443	0,0395	0,0142	0,0128
n=20				
1	0,046	0,050	0,004	0,020
2	0,055	0,052	0,013	0,028
3	0,050	0,050	0,004	0,022
4	0,054	0,054	0,011	0,010
5	0,047	0,048	0,000	0,016
6	0,044	0,051	0,020	0,019
7	0,058	0,055	0,010	0,016
8	0,041	0,053	0,010	0,022
9	0,047	0,045	0,016	0,014
10	0,059	0,056	0,017	0,003
11	0,042	0,052	0,004	0,030
12	0,045	0,037	0,004	0,009
13	0,045	0,053	0,000	0,020
14	0,040	0,054	0,014	0,023
15	0,048	0,052	0,012	0,018
16	0,049	0,053	0,020	0,020
17	0,057	0,045	0,040	0,030
18	0,043	0,051	0,010	0,000
19	0,046	0,057	0,000	0,018
20	0,049	0,052	0,018	0,022
Mean	0,04825	0,051	0,01135	0,018

We now tackle the performance of the second test. To do so, we need to generate data under the null hypothesis of non-increasing returns to scale (*nirs*). We consider separately the case of one input and two inputs.

In the one input case, given that the $\left(x_{ij}\right)$ are generated as uniform *iid* random variables on the interval $[1,9]$, the output of the j^{th} DMU can be generated as follows:

$$y_j = \begin{cases} x_{1j} e^{-0.1|v_j|} & \text{if } x_{1j} \le 5 \\ \left(\frac{5}{2} + \frac{1}{2} x_{1j}\right) e^{-0.1|v_j|} & \text{otherwise} \end{cases} \tag{17}$$

In the two-input case, we split the returns to scale regions not based on input levels, as in the one input case, but based on a reference output level. The procedure is as follows. First, using the randomly generated inputs, we compute the reference output level for the j^{th} DMU, $\bar{y} = x_{1j}^{1/2} x_{2j}^{1/2}$. Then the output of this DMU is generated according to the following formulae:

$$y_j = \begin{cases} x_{1j}^{1/2} x_{2j}^{1/2} e^{-0.1|v_j|} & \text{if } \bar{y} \le 5 \\ \sqrt{5} x_{1j}^{1/4} x_{2j}^{1/4} e^{-0.1|v_j|} & \text{otherwise} \end{cases} \tag{18}$$

This insures that all input combinations on the same isoquants exhibit the same type of returns to scale. To evaluate the performance of the second test, we proceed as we did with test #1, and we conduct the same eight Monte-Carlo experiments (the combinations of *s*=1, *m*=1, 2, □ =0,01, 0,05, and *n*=10, 20). The results of these experiments are summarized in Table 2. The results are very similar to those of Table 1. The simulated sizes are very close to the nominal sizes, again confirming the good performance of our test. When the number of DMU is set to ten, *n*=10, and $m = s = 1$, the average simulated size is equal to 0,047, a distortion of 0,003 for a nominal test size of 0,05. When the sample of DMU is increased to twenty, $n = 20$, the distortion of the size is even smaller, as it is equal to 0,0023. The distortion is zero for DMUs number 2, 4 and 16. The distortion increases however when the number of inputs increases to 2 from 1 (all other parameters equal). As in the case of the first test, this may be attributed to the curse of dimensionality. The same phenomena are observed at size $\alpha = 0,01$. For $m = s = 1$ and $n = 10$, the simulated size is 0,0155 and decreases to 0,01395 when the number of DMUs increases to twenty ($n = 20$). The difference between the nominal size and the simulated size is null or virtually null for three DMUs (4, 6 and 19). Again, these results confirm our initial conclusions that the performance of the test is good and improves when the size of the sample increases, as expected.

Table 2: Monte-Carlo estimates of size
Test #2 (H_0 : NIRS)

DMU	Nominal size			
	$\alpha = 0,05$		$\alpha = 0,01$	
	$m = s = 1$	$m = 2, s = 1$	$m = s = 1$	$m = 2, s = 1$
n=10				
1	0,047	0,060	0,017	0,028
2	0,043	0,036	0,012	0,012
3	0,050	0,041	0,018	0,019
4	0,048	0,039	0,018	0,014
5	0,057	0,039	0,023	0,015
6	0,053	0,044	0,017	0,017
7	0,047	0,028	0,015	0,004
8	0,041	0,047	0,011	0,014
9	0,045	0,035	0,015	0,007
10	0,039	0,057	0,009	0,020
Mean	0,047	0,0426	0,0155	0,015
n=20				
1	0,058	0,057	0,008	0,018
2	0,050	0,052	0,017	0,008
3	0,064	0,120	0,014	0,024
4	0,049	0,031	0,012	0,010
5	0,057	0,068	0,020	0,013
6	0,052	0,070	0,011	0,016
7	0,048	0,059	0,014	0,018
8	0,046	0,044	0,016	0,005
9	0,054	0,056	0,014	0,016
10	0,038	0,050	0,012	0,022
11	0,090	0,062	0,016	0,020
12	0,053	0,062	0,023	0,014
13	0,044	0,069	0,018	0,026
14	0,046	0,055	0,007	0,021
15	0,038	0,038	0,016	0,018
16	0,051	0,057	0,011	0,019
17	0,044	0,036	0,006	0,012
18	0,058	0,024	0,018	0,010
19	0,052	0,030	0,010	0,018
20	0,054	0,042	0,016	0,009
Mean	0,0523	0,0541	0,01395	0,01585

The performance of a test is not only related to its correct size, it must also have a high power. In the following we construct Monte-Carlo experiments to assess the power of our test. The general framework of the experiments is as follows. Data are generated under the alternative hypothesis and used to test for constant returns to scale. The power of the test is estimated by reporting the ratio of the number of time the null hypothesis is rejected over the number of Monte-Carlo iterations (*N*) for a given nominal size.

This experiment is repeated by increasing the divergence of the alternative hypothesis away from the null.

The experiments are performed for a one-input, one-output technology $(m=s=1)$ at nominal size $\alpha=0,05$ for a sample of twenty DMU, *n*=10. We suppose that the inputs (x_{1j}) are generated according to a uniform distribution on the interval $\left[-5\left[1-\tan\left(\frac{3\pi}{4}-\frac{\varphi}{2}\right)\right]/\tan\left(\frac{3\pi}{4}-\frac{\varphi}{2}\right),9\right]$ where $\varphi\in\left]\frac{\pi}{2},\pi\right]$. To generate the outputs (y_j), we first return an *iid* standard normal random variables v_j and then we apply the following transformations:

$$y_j=\left\{5\left[1-\tan\left(\frac{3\pi}{4}-\frac{\varphi}{2}\right)\right]+\tan\left(\frac{3\pi}{4}-\frac{\varphi}{2}\right)x_{1j}\right\}e^{-0.1|v_j|} \quad \text{if } x_{1j}\in\left[-5\left[1-\tan\left(\frac{3\pi}{4}-\frac{\varphi}{2}\right)\right]/\tan\left(\frac{3\pi}{4}-\frac{\varphi}{2}\right),5\right[\tag{19}$$

and

$$y_j=\left\{5\left[1-\tan\left(\frac{\varphi}{2}-\frac{\pi}{4}\right)\right]+\tan\left(\frac{\varphi}{2}-\frac{\pi}{4}\right)x_{1j}\right\}e^{-0.1|v_j|} \quad \text{if } x_{1j}\in[5,9] \tag{20}$$

The data generating process defined by equations (19) and (20) is a piecewise linear functions with the two segments intersecting at the point (5,5). On the first part of the function (equation (20)), the returns to scale are increasing while the technology exhibits decreasing returns to scale on the second part (given by equation (20)). We also add a further condition on the data generating process. All input-output combinations, (x_{1j},y_j), have the same chance of being generated under the *irs* technology (that is, for values of *x* smaller than five) as under the *drs* technology (that is, for values of *x* larger than five).

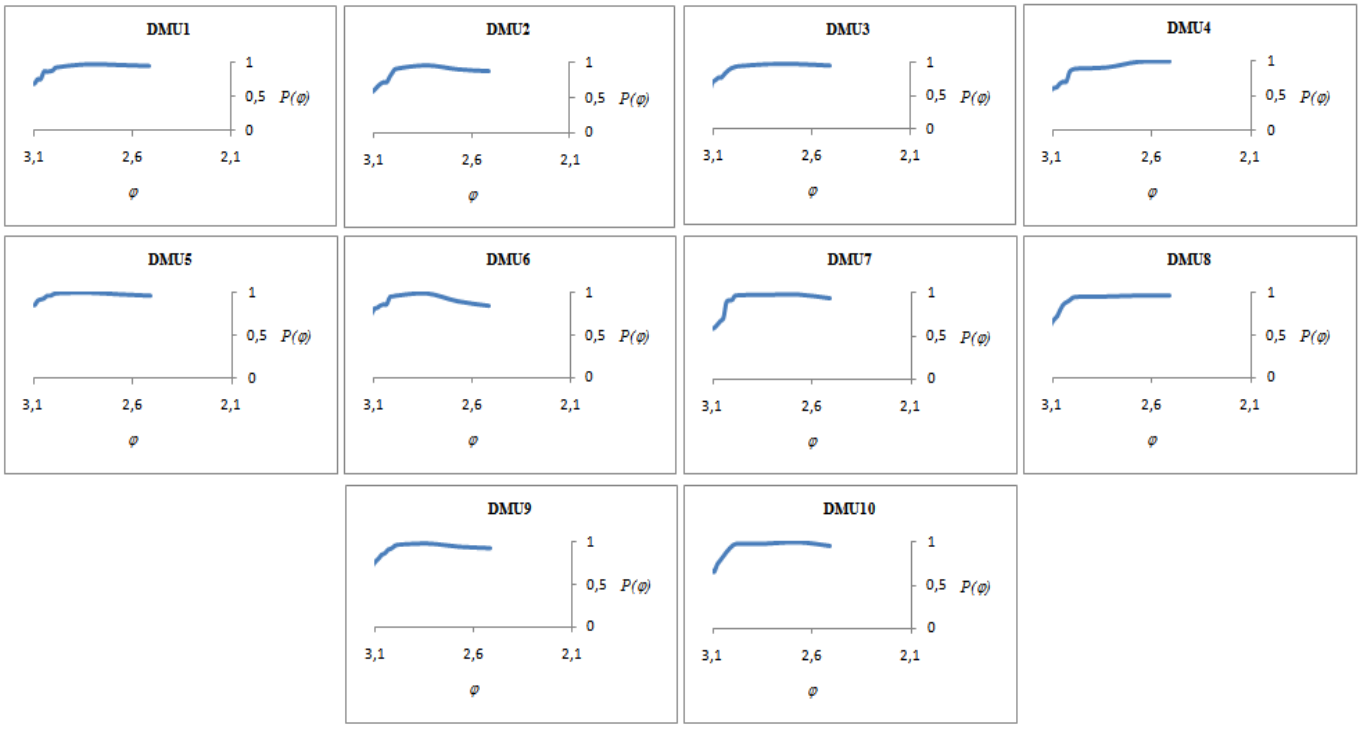
When $\varphi=\pi$, equations (19) and (20) simplifies to $y_j=x_je^{-0.1|v_j|}$, which is a constant return to scale technology. The angle formed by the ray from the origin through (5,5) (the *crs* technology) and either of the two segments of the technology is equal to $\pi-\varphi/2$. The two segments of the technology form a right angle at (5,5) when $\varphi=\pi/2$. It is important to control for the type of returns to scale during the course of the Monte-Carlo experiment. This is done by changing the value assumed by φ; this controls for the degrees of departure from the null hypothesis (i.e., the technology is assumed to exhibit constant returns to scale). Increasing φ increases the magnitude of the kink and pushes the technology away from the *crs* technology.

To perform the Monte-Carlo experiments using the DGP defined by equations (19) and (20), we set the following values for the angle $\varphi\in\left\{\pi-k\frac{\pi}{2}\,|\,k=0,01,...0,09,\ldots\ 0,1,\ldots 0,4\right\}$ and we construct the power function, $P(\varphi)$, for each DMU for test #1 when *n*=10. Again, the Monte-Carlo experiments consists in *B*=2000 bootstrap simulations for *N*=1000 Monte-Carlo iterations for each DMU.

The results of the simulation experiments are represented on Figure 1 (while the numbers used to compute theses figures are in Table A1 of the Appendix). On average, the value of the power function is 0.5696 for $\varphi=3,125$ (i.e. *k*=0,01) and 0,6714 for $\varphi=3,11$ (i.e. *k*=0,02). The power function increases continuously up to *k*=0,2 ($\varphi=2,827$) and continues to increase for many DMUs after. However, in some cases, the power decreases slightly when k increases. The average value of the peaks is 0,9746. After the peak, the power function seems to exhibit a very slow decrease (0,9639 for *k*=0,3 and 0,9407 for *k*=0,4). This phenomenon is generalized to all DMUs. For example, for DMU number one, we have a value of 0,593 for *k*=0,01 and a peak at 0,987 for *k*=0,2.

The highest power is reached by DMU 6 and assumes the value 0,996 for *k*=0,2. DMU 8 has power equal to 0,972 for values of *k*=0,3 and 0,4. In general, at *k*=0,2 and above, the power peaks and remains fairly close to one.

Figure 1: Power Functions $(\alpha = 0,05,\ n = 10,\ m = s = 1)$



The decreasing power function for high values of *k* is most likely related to the two-piece linear technology. The decrease in the power function happens in the neighbourhood of angle values $\varphi = 2,827$ (i.e. *k*=0,2) and $\varphi = 2,513$ (i.e. *k*=0,4). The contraction of the support of the uniform distribution that generated the inputs as a result of the larger angles leads to many input-output combinations close to the point (5,5) where the technology exhibits constant returns to scale. Even though the power function does not reach the value one, the power is inversely related to the angle, so as the data generating process drifts away from the null hypothesis, the power increases. This unexpected result for the shape of the power curve can be explained by the very small dimension of technology $(m = s = 1)$. Moreover, showing estimated size and especially power for only single output and single input case is misleading, in the light of the "curse of dimensionality" problem. However, with a complex computation it is difficult to make simulations for a technology for more than one input and one output.

4. Conclusion

In this article we have developed a procedure to test non parametric statistical hypothesis concerning the scale efficiency of organization. Because there is no asymptotic distribution for the test statistic under the null hypothesis, we used a smooth bootstrap methodology to approximate it. This approach allows us to estimate the *p*-values of the test and to determine a decision rule to accept or reject the null hypothesis. Our approach is based on the work of Simar and Wilson (2002) which use smooth bootstrap methodology to solve the hypothesis test problem. However in this paper, we have developed an individual test procedure not global as them. This procedure allows us to determine the nature of the scale efficiency of each organization.

To assess the performance of our test, we perform limited Monte-Carlo experiments. We chose to limit the number of cases under study because of the complexity of the calculations involved (the number of experiments for each test requires to solve $(N \times n \times (B+1))$ linear programs, and obviously the computing time increases with the number of sample points, *n*). The first step of the experiments was to contrast the nominal and real size of the test. In general, the distortion between the nominal size and the simulated value is very small.

When the sample size increases, this distortion decreases substantially, but without surprises increases when the dimension of the problem increases (e.g. when the number of inputs is larger, as it is the case in our experiments). This latter result can be attributed to the curse of dimensionality of nonparametric problems. The second step of the experiments was to assess the power of the test. In a limited scale problem (one-input and one-output), using a piecewise linear frontier, we found an unexpected form for the power curves: start at about 0.5 or even higher and are quite flat, often not reaching unity and sometimes falling. This can be explained by the problem of "curse of dimensionality". Indeed, showing estimated size and especially power for only single output single input case is misleading, in the light of the "curse of dimensionality" problem. The study of the performance of our test requires simulations with higher dimension technology. This complicates the computation and requests a prohibitive execution time.

References

Banker, R.D (1993). Maximum likelihood, consistency and data envelopment analysis: a statistical foundation. Management Science 39, 1265-1273.

Banker, R.D (1996). Hypothesis tests using data envelopment analysis. Journal of Productivity Analysis 7, 139-159.

Farrell, M. J. (1957). The Measurement of Productive Efficiency. Journal of the Royal Statistical Society, Series A, 120 (3), 253-290.

Färe, R. (1988). Fundamentals of Production Theory. Lecture Notes in Economics and Mathematical Systems, Vol. 311, Springer-Verlag, Berlin Heidelberg.

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

Jeong, S.O (2004). Asymptotic distribution of DEA efficiency scores. J.Korean Stat.Soc. 33, 449-458.

Kittelsen, S.A.C. (1997). Monte Carlo Simulations of DEA Efficiency Measures and Hypothesis Tests. *Unpublished Working Paper*, Gaustadalleén 21, N-0371 Oslo, Norway.

Kneip, A., Park, B.U. & Simar, L. (1998). A Note on the Convergence of Nonparametric DEA Estimators for Production Efficiency Scores. Econometric Theory 14, 783-793.

Kneip, A., Simar, L. & Wilson, P.W. (2009). A Computationally efficient, consistent bootstrap for inference with non-parametric DEA estimators. *Discussion Paper* No. 0903, Institut de Statistique, Université Catholique de Louvain, Louvain-la-Neuve, Belgium.

Kneip, A., Simar, L. & Wilson, P. W. (2008). Asymptotics and consistent bootstraps for DEA estimators in non-parametric frontier models. Econometric Theory, 24, 1663-1697.

Löthgren, M. & Tambour, M. (1999). Testing Scale Efficiency in DEA Models: a Bootstrapping Approach. Applied Economics, 1999, 31, 1231-1237.

Park B.U, Jeong S-O & Simar L. (2010). Asymptotic distribution of conical-hull estimators of directional edges. Ann Stat 38, 1320-1340.

Schuster, E.F. (1985). Incorporating Support Constraints into Nonparametric Estimators of Densities. Communications in Statistics-Theory and Method 14, 1123-1136.

Shephard, R.W. (1970). Theory of Cost and Production Functions. Princeton, N.J.: Princeton University Press.

Silverman, B.W. (1986). Density Estimation for Statistics and Data Analysis. New York: Chapman and Hall.

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

Simar, L. and Wilson, P.W. (1999). Some Problems with the Ferrier/Hirschberg Bootstrap Idea. Journal of productivity Analysis 11, 67-80.

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

Simar L, Wilson, P.W (2000b). A general methodology for bootstrapping in non-parametric frontier models. J. Appl. Stat. 27, 779-802.

Simar, L. and Wilson, P.W. (2002). Non-parametric tests of returns to scale. European Journal of Operational Research 139, 115-132.

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

Appendix

Table A1 : Power Functions

K	DMU1	DMU2	DMU3	DMU4	DMU5	DMU6	DMU7	DMU8	DMU9	DMU10
0,01	0,593	0,515	0,541	0,502	0,671	0,684	0,528	0,502	0,606	0,554
0,02	0,684	0,567	0,697	0,58	0,84	0,753	0,58	0,619	0,723	0,671
0,03	0,697	0,606	0,736	0,619	0,853	0,818	0,593	0,684	0,775	0,649
0,04	0,762	0,658	0,775	0,632	0,905	0,831	0,632	0,723	0,812	0,738
0,05	0,762	0,697	0,775	0,684	0,918	0,857	0,671	0,796	0,855	0,784
0,06	0,875	0,723	0,832	0,71	0,931	0,87	0,71	0,863	0,871	0,834
0,07	0,875	0,723	0,877	0,71	0,962	0,87	0,897	0,892	0,914	0,884
0,08	0,882	0,81	0,912	0,846	0,962	0,953	0,91	0,915	0,927	0,927
0,09	0,896	0,874	0,927	0,883	0,98	0,961	0,923	0,945	0,954	0,959
0,1	0,938	0,921	0,943	0,896	0,991	0,97	0,968	0,956	0,972	0,976
0,2	0,987	0,973	0,975	0,914	0,998	0,996	0,975	0,961	0,986	0,981
0,3	0,976	0,914	0,975	0,993	0,979	0,905	0,979	0,972	0,949	0,997
0,4	0,961	0,887	0,954	0,997	0,962	0,853	0,936	0,972	0,931	0,954