# How to Bootstrap DEA Estimators: A Monte Carlo Comparison*

Mickael Löthgren
Department of Economic Statistics
Stockholm School of Economics
P.O. Box 6501, S-113 83 Stockholm, Sweden.
E-mail: stml@hhs.se.
Fax: +46 - 8 - 34 81 61.

*Working Paper Series in Economics and Finance, No. 223*
*February 1998*

**Abstract**

This paper evaluates the performance of three bootstrap algorithms for the data envelopment analysis (DEA) estimator using a Monte Carlo simulation study. The Löthgren and Tambour (1997) (LT) algorithm; the Simar and Wilson (1997b) (SW) algorithm; and a combination of the LT and SW algorithms (the LSW-algorithm) are considered in the study. The empirical coverage accuracy of bootstrap confidence intervals are simulated under both variable returns to scale (VRS) and constant return-to-scale (CRS) restricted DEA estimators. The results indicate that the LSW-algorithm performs slightly better than the LT-algorithm, which in turn performs better than the SW-algorithm.

**Key Words:** Bootstrap; Confidence Intervals; Data Envelopment Analysis; Monte Carlo Simulation.
**JEL-Classification:** C14; C15; D24.

---

*I thank Mickael Salabasis for helpful discussions of the nonparametric smoothing procedure. Financial support from the Jan Wallander Foundation for Research in the Social Sciences is gratefully acknowledged.

## 1. Introduction

Estimation of technical efficiency using nonparametric piecewise linear envelopment of the data dates back to the work of Farrell (1957). The data envelopment analysis (DEA) estimator of technical efficiency, introduced by Charnes, Cooper and Rhodes (1978), have found use in many efficiency measurement studies and have been modified and extended in several ways as presented in the DEA survey by Seiford (1996). Recently the statistical properties of the DEA estimators have been investigated by several authors. Banker (1993), establishes consistency of the DEA estimators for the single-output, multiple-input case, and Kneip, Park and Simar (1996) provide convergence results for the general multiple-input, multiple-output case. Korostelev, Simar and Tsybakov (1995a, 1995b) establish the consistency of the FDH and DEA production set estimators and their convergence rates.

An important issue in DEA applications is whether a given set of finite sample DEA efficiency estimates indicate significant technical inefficiencies of the evaluated firms or not. This information is reasonably of importance to managers who wish to focus their efforts to increase efficiency where improvements are most needed. Albeit consistent, real world applications of DEA estimators offer no guidance to the statistical inference problem in that only point estimates of efficiency are obtained from the estimators.

An approach to perform the desired inference is to bootstrap the DEA estimator. The bootstrap is a computationally intensive general method that relies on a simple idea to perform statistical inference in various problems where the analytical properties of the considered estimator is unknown or hard (or even impossible) to derive analytically. The idea in the bootstrap method, introduced by Efron (1979) and later extended in several directions, as presented in, e.g., Efron and Tibshirani (1993), is to resample the estimator and use the empirical distribution of the resampled estimates to calculate bootstrap confidence intervals or to construct bootstrap hypotheses tests to perform the desired statistical inference.

The application of the bootstrap to DEA estimators is rather recent and under extensive development. As of yet a consensus on a single bootstrap methodology for the DEA estimators remains to be developed. Several papers applying various bootstrap techniques on DEA estimators have recently appeared. See, e.g., Gstach (1995), Ferrier and Hirshberg (1997), Löthgren and Tambour (1997) and Simar and Wilson (1997b). Recently, Simar and Wilson (1997c) criticize the bootstrap procedure suggested by Ferrier and Hirshberg (1997) and present "a primer on bootstrapping nonparametric models" with a general guidance of how the DEA estimators should (and should not!) be bootstrapped.

The purpose of this paper is to take up the discussion in Simar and Wilson (1997c) and present some Monte Carlo simulation results in favor of an alternative simple DEA bootstrap procedure that offers a simple alternative to the algorithm suggested by Simar and Wilson (1997b). Specifically the bootstrap algorithm by Löthgren and Tambour (1997) (LT) is presented and compared to the algorithm by Simar and Wilson (1997b) (SW).

The LT-algorithm is a computationally simple DEA bootstrap algorithm that differs from the SW-algorithm in two important ways: First, a simple resampling of original efficiency estimates are used, without application of the smoothed resampling applied by SW. Second, a fundamental difference is that the calculation of the resampled efficiency estimates differ. In the LT-algorithm the bootstrap efficiency estimates are based on the distance from the pseudo-data to the bootstrap frontier estimates. This is in contrast to the SW-algorithm where the bootstrap efficiency estimates are based on the distance from the original observed inputs to the bootstrapped production frontier estimates.

A comparative Monte Carlo simulation study is conducted to investigate the performance of the LT and the SW algorithms to bootstrap single-period DEA estimators. A comparison of the relative performance of the LT and the SW algorithms is of interest by itself but does not provide any information on the possible causes of a possible difference in performance. To give some guidance on this issue we propose and evaluate the performance of a third algorithm (the LSW-algorithm) that is based on the smoothing procedure in the SW-algorithm and the use of the pseudo input data in the definition of the bootstrap efficiency estimator as in the LT-algorithm. A comparison of the relative performance of the LT and LSW algorithms indicates whether any differences in performance is due to the use of a smoothed (SW) or non-smoothed (LT) resampling of pseudo-efficiencies in the bootstrap algorithm. Furthermore, a comparison of the relative performance of the SW and LSW algorithms indicates whether any differences in performance is due to the use of original data (SW) or pseudo-data (LT) in the evaluation of the bootstrap efficiency estimates.

The Monte Carlo study considers estimation of input-oriented[1] technical inefficiency using a CRS technology as the true data generating process. Technical inefficiency is estimated using both the VRS and the CRS restricted DEA estima-

[1] This is not restrictive and the described algorithms and techniques can be trivially extended to an output-oriented setting.

tor. The coverage accuracy performance of simple percentile and bias-corrected bootstrap confidence intervals are considered for each of the LT, the SW and the LSW algorithms. For the computationally demanding VRS estimator we consider small-sample properties for sample sizes up to $n = 60$. For the CRS estimator we consider sample sizes up to $n = 500$ to get some indications of the consistency properties of the different bootstrap algorithms.

For the VRS estimator the results indicate that the combined LSW-algorithm performs slightly better than the LT-algorithm, which in turn performs better than the more computationally demanding SW-algorithm. For larger sample sizes and higher nominal confidence levels the difference in performance between the LT and SW-algorithms is more clear. For the CRS estimator the LSW-algorithm also shows the best performance. For all considered sample sizes, the LSW percentile intervals have consistently the highest coverage accuracies. The LT-algorithm intervals performs overall better than the SW-algorithm. As for the consistency properties of the algorithms the coverage accuracy of bootstrap confidence intervals from the LT and LSW algorithms show a clear convergence of the empirical coverage accuracy to the nominal levels for sample sizes larger than 120. For the SW-algorithm, on the other hand, the results cast doubts on the consistency properties of the algorithm.

The paper unfolds as follows: Section 2 presents the DEA estimator. Section 3 presents the bootstrap algorithms and Section 4 presents the bootstrap confidence intervals considered in the study. Section 5 presents the Monte Carlo simulation study and Section 6 concludes.

## 2. The DEA Estimator

A multiple-input, multiple-output production technology, where inputs $x \in R^d_+$ inputs are used in the production of $y \in R^p_+$ outputs can be represented by the production set $\Psi$ of attainable input-output combinations:

$$\Psi = \left\{ (x, y) \in R^{d+p}_+ : x \text{ can produce } y \right\}. \tag{2.1}$$

The technology is assumed to satisfy a set of axioms discussed in, e.g., Shephard (1953) and Shephard (1970). In short: (i) inactivity is allowed; (ii) “free lunch“ is not allowed; (iii) strong disposability of inputs and outputs; and (iv) the technology is convex.

An alternative, equivalent, representation of the technology is the input set, defined as

$$L(y) = \{x : (x, y) \in \Psi\}. \tag{2.2}$$

The production technology is completely characterized by the Farrell (1957) scalar-values input-based technical efficiency measure defined as

$$\theta(x, y) = \min\{\theta : \theta x \in L(y)\}. \tag{2.3}$$

This measure is reciprocal to the input distance function by Shephard (1953) and Shephard (1970). $\theta(x, y) \leq 1$ if and only if $x \in L(y)$. The value of the efficiency measure is given by $\theta(x, y) = \|x\| / \left\|x^f\right\|$, where $x^f \in IsoqL(y) = \{x : x \in L(y), \mu x \notin L(y), \mu < 1\}$ is the frontier input. A firm is considered as technically efficient if the efficiency measure equals one.

Following Charnes et al. (1978) (CCR) the variable returns to scale (VRS) DEA efficiency estimator for firm $i = 1, ..., n$, based on a sample $\chi_n = \{(x_i, y_i);\ i = 1, ..., n\}$, is given by the solution to the linear programs

$$\widehat{\theta}_{in}^{VRS} = \min\left\{\theta : \theta x_i \in \widehat{L}_n^{VRS}(y_i)\right\}, \tag{2.4}$$

where $x_i$ is the $d$-vector of inputs and $y_i$ is the $p$-vector of outputs for firm $i$, $\widehat{L}_n^{VRS}(y_i)$ is the piece-wise linear convex hull envelopment of the observed sample $\chi_n$ given by

$$\widehat{L}_n^{VRS}(y_i) = \left\{x : y_i \leq Yz,\ x \geq Xz,\ \sum_{i=1}^{n} z_i = 1,\ z \in R_+^n\right\}, \tag{2.5}$$

where $Y = (y_1, y_2, ..., y_n)$ is a $(p \times n)$ matrix, $X = (x_1, x_2, ..., x_n)$ is a $(d \times n)$ matrix of inputs and $z$ is a $n$-vector of non-negative intensity variables.

The constant returns to scale (CRS) efficiency estimator, $\widehat{\theta}_i^{CRS}$, is given by

$$\widehat{\theta}_{in}^{CRS} = \min\left\{\theta : \theta x_i \in \widehat{L}_n^{CRS}(y_i)\right\}, \tag{2.6}$$

where $\widehat{L}_n^{CRS}(y_i)$ is the piece-wise linear conical hull of the data, defined by

$$\widehat{L}_n^{CRS}(y_i) = \left\{x : y_i \leq Yz,\ x \geq Xz,\ z \in R_+^n\right\}. \tag{2.7}$$

As discussed in Simar and Wilson (1996), the "safest" approach in estimating efficiency, that avoids a possible misspecification, is to use the VRS estimator. This suggestion is motivated by the fact that for a CRS technology, both $\widehat{L}_n^{CRS} \subseteq L$

and $\widehat{L}_n^{VRS} \subseteq L$, and both input set estimators converge to the true input set $L$ as the sample size $n \rightarrow \infty$ (see, e.g., Korostelev, Simar and Tsybakov (1995b)). Hence, under CRS, both estimators $\widehat{\theta}_{in}^{VRS}$ and $\widehat{\theta}_{in}^{CRS}$ are consistent. If, on the other hand, the true technology does *not* exhibit CRS, the CRS input set estimator $\widehat{L}_n^{CRS}$ is misspecified and will not converge to $L$. This in turn implies that the CRS DEA estimator $\widehat{\theta}_{in}^{CRS}$ are inconsistent in this setting. The VRS estimator remains consistent, however. Kneip et al. (1996) provide a complete and rigorous treatment of the VRS DEA estimators consistency properties.

Albeit consistent, the estimators are biased in small-samples. Specifically, if the true technology exhibits CRS, the estimators are nested as $1 \leq \theta_i \leq \widehat{\theta}_{in}^{CRS} \leq \widehat{\theta}_{in}^{VRS}$, which implies that both the CRS and the VRS estimators are positively biased in this setting. If the true technology does *not* exhibit CRS, the VRS estimator remains consistent and positively biased for small samples whereas the CRS estimator is misspecified and inconsistent.

## 3. Bootstrapping the DEA Estimator

The bootstrap method is a well-established statistical resampling method used to perform inference in complex problems. The crucial step in any application of the bootstrap is a clear specification of the data generating process (DGP) underlying the observed data. The basic idea of the bootstrap method is to approximate the sampling distributions of the estimator by using the empirical distribution of resampled estimates obtained from a Monte Carlo resampling simulation of the estimation procedure where repeated resamples obtained from an estimate of the DGP produce repeated estimates. The performance of the bootstrap in terms of the validity of the conducted statistical inference ultimately hinges on how well the DGP characterizes the true data generation and how well the DGP is mimicked in the resampling simulation.

The DEA bootstrap algorithms by both Löthgren and Tambour (1997)[2] (LT) and Simar and Wilson (1997b) (SW) is based on the same DGP model where the inputs are assumed given by random radial deviations off the isoquant of the input set. Formally, each input in the sample of input-output observations

[2] The specified DGP in LT is output-oriented but is easily adapted to the input-based model considered here.

$\chi_n = \{(x_i, y_i) \, ; \; i = 1, ..., n\}$, are specified as

$$(x_i, y_i) = \left( x_i^f / \theta_i, y_i \right), \tag{3.1}$$

where $x_i^f \in IsoqL(y_i)$ is the unobservable frontier input for firm $i$. The true efficiency measures are assumed to be drawn from the same distribution, i.e., $\theta_i \sim F_\theta$, $i = 1, ..., n$.

Note that this DGP model represents the idea that, conditioned on the outputs and the input proportions, the stochastic elements in the production process are completely represented by the random input efficiency measures.

The main idea in the bootstrap simulation is to mimic the DGP. The procedure in both the LT and SW algorithms in each resample is as follows: Conditioned on observed outputs and input proportions, the resample data are constructed in two steps. First, the frontier inputs are estimated and bootstrap pseudo-inputs are generated by replicating the DGP in (3.1) using the estimated frontier inputs and pseudo efficiencies drawn from some estimate of the distribution $F_\theta$. The algorithm by LT is based on a simple resampling from the empirical distribution of the estimated efficiencies are used to generate the pseudo-efficiencies. The SW algorithm make use of a smoothed resampling procedure, based on consistency argument. Second, the bootstrapped efficiency estimate is obtained by evaluating the distance from the pseudo input (in the LT-algorithm) or the original input (in the SW-algorithm) relative to the bootstrap estimate of the frontier.

In the two subsections that follow we present the two algorithms, and discuss the differences between them, in detail.[3]

### 3.1. The LT-algorithm

The bootstrap algorithm proposed by LT is given by the following steps:

1. Transform the input-output vectors using the original efficiency estimates $\left\{ \widehat{\theta}_{in}, i = 1, ..., n \right\}$ as
$$\left( \widehat{x}_i^f, y_i, \right) = \left( x_i \cdot \widehat{\theta}_{in}, y_i \right). \tag{3.2}$$
2. Resample, independently with replacement, $n$ technical efficiencies from the set of original estimates $\left\{ \widehat{\theta}_{in} \right\}$. Let $\delta_i^*$, $i = 1, ..., n$, denote the resampled efficiencies.

[3]The algorithm can be applied to both the VRS and the CRS DEA estimator. For notational simplicity we suppress the scale restriction notation in the following and focus the presentation on the VRS estimator.

3. Let the bootstrap pseudo-data be given by

$$\left(x_i^*, y_i^*\right) = \left(\widehat{x}_i^f / \delta_i^*, y_i\right). \tag{3.3}$$

4. Estimate the bootstrap efficiencies using the pseudo-data and the linear program (2.5) as[4]

$$\widehat{\theta}_{in}^{LT*} = \min_{\theta,z} \left\{ \theta : y_i \leq Yz,\ \theta x_i^* \geq X^* z,\ \sum_{i=1}^{n} z_i = 1,\ z \in R_+^n \right\}. \tag{3.4}$$

5. Repeat steps (2) - (4) $B$ times to create a set of $B$ firm-specific bootstrapped efficiency estimates $\widehat{\theta}_{in}^{LT*b}$, $b = 1, ..., B$.

In this algorithm the bootstrap frontier estimates and the bootstrap efficiency estimates are resampled based on a resampling of technical efficiencies from the empirical distribution of the original estimates of efficiency. Furthermore, the bootstrap replicates of efficiency is based on the resampled data in the same manner as the original estimates are based on the original data.

### 3.2. The SW-algorithm

The bootstrap algorithm proposed by SW differs from the LT-algorithm in step 2 and step 4.

In step 2 a smoothing procedure, based on a kernel smoothing of the empirical distribution of the original efficiency estimate, is utilized to generate a smoothed resample of pseudo-efficiencies. The utilized smoothing procedure is based on the reflection modification of the Gaussian kernel density estimate discussed in Silverman (1986). The procedure is described in detail in SW, here we only present a basic discussion sufficient to implement the smoothing.

Let $\delta_i^*$ be the non-smoothed resample drawn independently, with replacement, from the empirical distribution of the original estimates of technical efficiency $\left\{\widehat{\theta}_{in}\right\}$. The smoothing procedure consists of two steps: First a small perturbation is added to $\delta_i^*$, and second a correction of the resampled sequence is applied.

First, a small perturbation $h\varepsilon_i^*$ is added to $\delta_i^*$, where $h$ denotes the bandwidth parameter and $\varepsilon_i^*$ is drawn *i.i.d.* from a standard normal distribution, to generate

[4] The CRS estimator is obtained by deleting the restriction $\sum_{i=1}^{n} z_i = 1$. $X^* = (x_1^*, x_2^*, ..., x_n^*)$ is a $(d \times n)$ matrix of pseuso-inputs.

the smoothed pseudo-efficiency $\widetilde{\delta}_i^*$. Due to the fact that the efficiency measure is bounded on the unit interval as discussed in Section 2, the following reflection procedure is utilized to generate $\widetilde{\delta}_i^*$:

$$\widetilde{\delta}_i^* = \begin{cases} \delta_i^* + h\varepsilon_i^* & \text{if } \delta_i^* + h\varepsilon_i^* \leq 1 \\ 2 - (\delta_i^* + h\varepsilon_i^*) & \text{otherwise.} \end{cases} \tag{3.5}$$

If the realization $\delta_i^* + h\varepsilon_i^* > 1$, $\widetilde{\delta}_i^*$ is set to the symmetric image of $\delta_i^* + h\varepsilon_i^*$ reflected in the boundary point, i.e., $\widetilde{\delta}_i^* = 2 - (\delta_i^* + h\varepsilon_i^*)$.

One important problem in the application of the smoothing procedure is the choice of the bandwidth parameter $h$. Several approaches to select the bandwidth $h$ is available, as discussed in, e.g., Silverman (1986). In the Monte Carlo study performed in this paper, we use a simple automatic robust bandwidth selection rule for univariate data proposed by Silverman (1986, p. 47-48)[5]

$$h = 0.90n^{-1/5} \min\left\{\widehat{\sigma}_{\widehat{\theta}},\ R_{13}/1.34\right\}, \tag{3.6}$$

where $\widehat{\sigma}_{\widehat{\theta}}$ denotes the plug-in standard deviation estimate of the efficiency estimates $\left\{\widehat{\theta}_{in}\right\}$ and $R_{13}$ denotes the interquartile range of the empirical distribution of $\left\{\widehat{\theta}_{in}\right\}$, respectively.

The final smoothed resampled efficiencies, denoted $\gamma_i^*$, are obtained by correcting $\widetilde{\delta}_i^*$ as

$$\gamma_i^* = \overline{\delta}_i^* + \left(\widetilde{\delta}_i^* - \overline{\delta}_i^*\right) / \sqrt{1 + h^2/\widehat{\sigma}_{\widehat{\theta}}^2}, \tag{3.7}$$

where $\overline{\delta}_i^* = \sum_{i=1}^n \delta_i^*/n$ is the average of the resampled original efficiencies. This correction guarantees that the resampled efficiency (asymptotically) have the same first two moments as the original efficiency estimates $\left\{\widehat{\theta}_{in}\right\}$.

The second, more fundamental difference between the two algorithms is in step 4. In the SW-algorithm the bootstrap efficiency estimate for the $i$th firm is evaluated as the efficiency of the original input $x_i$ relative to the bootstrapped isoquant of the input set $\widehat{L}^*(y_i)$ instead of using the bootstrap pseudo-input $x_i^*$ as in the LT-algorithm.

To summarize, the SW-algorithm is given by the following steps:

1. Transform the input-output vectors using the original efficiency estimates $\left\{\widehat{\theta}_{in}, i = 1, ..., n\right\}$ as

$$\left(\widehat{x}_i^f, y_i,\right) = \left(x_i \cdot \widehat{\theta}_{in}, y_i\right). \tag{3.8}$$

[5] A similar bandwidth selection rule for bivariate smoothing is used by Simar and Wilson (1996) for the smoothed bootstrap of Malmquist productvity indices.

2. Generate smoothed resampled pseudo-efficiencies $\gamma_i^*$ as follows

   2.1. Given the set of estimated efficiencies $\left\{\widehat{\theta}_{in}\right\}$, use (3.6) to obtain the bandwidth parameter $h$.

   2.2. Generate $\{\delta_i^*\}$ by resampling, with replacement, from the empirical distribution $\left\{\widehat{\theta}_{in}\right\}$ of the estimated efficiencies.

   2.3. Generate the sequence $\left\{\widetilde{\delta}_i^*\right\}$ using (3.5).

   2.4. Generate the smoothed pseudo-efficiencies $\{\gamma_i^*\}$ using (3.7).

3. Let the bootstrap pseudo-data be given by

$$(x_i^*, y_i^*) = \left(\widehat{x}_i^f / \gamma_i^*, y_i\right). \tag{3.9}$$

4. Estimate the bootstrap efficiencies using the pseudo-data and the linear program (2.5) as[6]

$$\widehat{\theta}_{in}^{SW*} = \min_{\theta,z} \left\{\theta : y_i \leq Yz,\ \theta x_i \geq X^* z,\ \sum_{i=1}^{n} z_i = 1,\ z \in R_+^n\right\}. \tag{3.10}$$

5. Repeat steps (2) - (4) $B$ times to create a set of $B$ firm-specific bootstrapped efficiency estimates $\widehat{\theta}_{in}^{SW*b}$, $i = 1, ..., n$, $b = 1, ..., B$.

### 3.3. The LSW-algorithm

We also consider a third algorithm which is based on a combination of the LT and the SW algorithms. The LSW-algorithm is based on the smoothing procedure in the SW-algorithm and the use of the pseudo input data in the definition of the bootstrap efficiency estimator as in the LT-algorithm. That is, step 2 in the combined LSW-algorithm consists of the smoothing procedure in step 2 in the SW-algorithm; and in step 4 in the LSW-algorithm the bootstrap replicate of the efficiency estimate is obtained according to step 4 in the LT-algorithm.

[6] The CRS estimator is obtained by deleting the restriction $\sum_{i=1}^{n} z_i = 1$.

### 3.4. Some remarks

SW argues that a simple naive resampling cannot give a consistent bootstrap procedure and concludes that the proposed application of the smoothing and reflection procedure in the resampling in the SW-algorithm provides a way to obtain a consistent bootstrap procedure. In their discussion of the consistency issue, SW do not consider any alternative approach to calculate the bootstrap efficiency estimates other than defining the bootstrap efficiency estimates as the efficiency of the original observed input relative to the bootstrap frontier estimate. SW did not consider the alternative approach used in the LT-algorithm where the bootstrap efficiency estimates are based on the distance from the pseudo input data (in the input-oriented efficiency approach considered in this paper) to the bootstrap frontier estimates. This difference between the algorithms is crucial since without this modification the simple resampling used in the LT-algorithm will give an inconsistent bootstrap procedure as discussed in a single-input example by Simar and Wilson (1997c).

As discussed by SW and Simar and Wilson (1997c) a key issue in any bootstrap application is the replication of the data generating process assumed to generate the observed data. We argue that the proposed LT-algorithm provides a more accurate replication of the DGP specified in (3.1) than the SW-algorithm. While the resampling suggested by SW focuses on the sampling variability of the production frontier estimate underlying the efficiency estimate, the resampling in the LT-algorithm explicitly incorporates the fact that the sampling distribution of the DEA estimator necessarily is a function of *both* the sampling distribution of the production frontier estimate and the sampling distribution of the input-output data. In the LT-algorithm the bootstrap frontier estimates and the bootstrap efficiency estimates are based on the resampled data in the same manner as the original estimates are based on the original data. In this way both the stochastic data generating process and the sampling distribution of the frontier estimates will be incorporated in the bootstrapped efficiency estimates.

## 4. Confidence Intervals

### 4.1. Simple percentile intervals

The simplest and most straightforward method to obtain bootstrap confidence intervals is the percentile method. As presented in, e.g., Efron and Tibshirani (1993), the percentile interval is based on the empirical distribution function of the bootstrapped efficiencies $\widehat{\theta}_{in}^{*b}$, $b = 1, ..., B$, defined as $\widehat{F}_i(s) = \frac{1}{B}\sum_{b=1}^{B} I\left(\widehat{\theta}_{in}^{*b} \leq s\right)$, for any real value $s$, where $I(\cdot)$ denotes the standard indicator function. A $(1-2\alpha)$ (equal-tail) confidence interval for the true technical efficiency $\theta$ for the $i$th firm is given by

$$\left(\widehat{\theta}_{in}^{*(\alpha)},\ \widehat{\theta}_{in}^{*(1-\alpha)}\right), \tag{4.1}$$

where $\widehat{\theta}_{in}^{*(\alpha)}$ is the $\alpha$th quantile of $\widehat{F}_i$, i.e., $\widehat{\theta}_{in}^{*(\alpha)} = \widehat{F}_i^{-1}(\alpha)$. The quantiles of $\widehat{F}_i$are given by the $[(B+1)\alpha]$th and the $[(B+1)(1-\alpha)]$th ordered values of $\widehat{\theta}_{in}^{*b}$, $b = 1, ..., B$, respectively, where $[r]$ denotes the integer part of any real value $r$.

### 4.2. Bias-corrected intervals

As discussed above in Section 2 the DEA estimators are biased in small-samples. Simar and Wilson (1997b) present a simple and direct approach to bias correct the percentile intervals in (4.1) using a simple additive bias correction. The bootstrap estimate of the DEA estimator bias is given by

$$\widehat{bias}_i^* = \frac{1}{B}\sum_{b=1}^{B}\widehat{\theta}_{in}^{*b} - \widehat{\theta}_{in}. \tag{4.2}$$

The bias-corrected firm-specific $(1-2\alpha)$ (equal-tail) confidence intervals are simply obtained by shifting the bounds in the intervals in (4.1) by the factor $2 \cdot \widehat{bias}_i^*$ as

$$\left(\widehat{\theta}_{in}^{*(\alpha)} - 2 \cdot \widehat{bias}_i^*,\ \widehat{\theta}_{in}^{*(1-\alpha)} - 2 \cdot \widehat{bias}_i^*\right). \tag{4.3}$$

Simar and Wilson (1997b) motivate the correction of $2 \cdot \widehat{bias}_i^*$ by the fact that this correction centers the empirical bootstrap distribution on the bias-corrected estimate $\widetilde{\theta}_{in} = \widehat{\theta}_{in} - \widehat{bias}_i^*$. Shifting the intervals in (4.3) by a factor of $1 \cdot \widehat{bias}_i^*$ will center the intervals on the biased estimator $\widehat{\theta}_{in}$.

## 5. The Monte Carlo Study

The simulation study consider a single-input, single-output technology with DGP specified in accordance with assumptions *A1* to *A4* in Kneip et al. (1996) used to establish the consistency properties of the DEA estimator. In short, these assumptions are[7]

1. The sample $\chi_n = \{(x_i, y_i); i = 1, ..., n\}$ are *i.i.d.* random variables on the convex attainable production set $\Psi$.

2. The support of the density function of the output $y$, $f(\cdot)$, has a compact support $\mathcal{D} \subseteq R_+$.

3. The inputs has a conditional density $f(x|y)$ for all $y \in \mathcal{D}$.

4. To ensure consistency of the DEA estimator, the DGP must insure that observations will be observed near the frontier when $n$ is sufficiently large.[8]

Based on these assumptions, the following single-input, single-output constant returns to scale DGP is considered in this study:

The outputs are assumed uniformly distributed on the compact support $\mathcal{D}$ given by the interval $\mathcal{D} = [1.0,\ 9.0]$. I.e., $y_i \sim U(1.0, 9.0)$. Given the output, the frontier inputs $x_i^f$ are derive from the simplest possible CRS technology relation $y_i = x_i^f$. The actually observed input for the $i$th firm is then generate by the following input-oriented technical inefficiency mechanism

$$x_i = x_i^f e^{0.10|u_i|}, \tag{5.1}$$

where $u_i$ is *i.i.d.* standard normally distributed.

Under this DGP, input-oriented technical efficiency is given by $\theta_i = e^{-U_i}$, where $U_i \sim N(0, 0.010)$ truncated at zero from below.

In each Monte Carlo replicate, input and output data are generated according to the design above for a sample of $n$ firms. For each firm in the sample, technical inefficiency is estimated using the linear programs presented in Section 2. The bootstrap algorithms presented in Section 3 gives $n$ bootstrap confidence intervals for the firm-specific technical inefficiency. For each replicate of the Monte Carlo simulation we keep track of whether or not the intervals cover the true realizations of the technical inefficiency. Completing this procedure, the empirical coverage of the Monte Carlo replicates give estimates of the true coverage accuracies, i.e., the empirical confidence level, of the firm-specific bootstrap confidence intervals.

[7] Note that the assumptions in Kneip et al. (1996) are stated for general multiple-input, multiple-output technologies.

[8] See assumption *A4* in Kneip et al. (1996) for a formal expression of this assumption.

In the simulation study, we use 1000 Monte Carlo replicates and $B = 1000$ bootstrap resamples in the bootstrap algorithm. This number of bootstrap replicates is recommended by Efron and Tibshirani (1993, p. 275) in order to make the variability of the boundaries of the confidence intervals constructed from the bootstrap "acceptably" low.

### 5.1. Results[9]

Results are presented for the empirical coverage accuracy of simple percentile and bias-corrected bootstrap confidence intervals for the VRS and the CRS estimator for nominal confidence levels 0.80, 0.90, 0.95 and 0.975, 0.99. For the VRS estimator we consider small-sample properties for sample sizes $n = 20$, 30 and 60. For the more computationally simple CRS estimator we also consider the additional sample sizes $n = 120$, 250 and 500 to assess some consistency properties of the coverage accuracy of the bootstrap confidence intervals obtained from the algorithms.

*Table 5.1* gives Monte Carlo estimates of the empirical coverage accuracy of the simple percentile and bias-corrected bootstrap confidence intervals obtained when bootstrapping the VRS DEA estimator. *Table 5.2* gives similar results for the CRS DEA estimator.[10]

For both the VRS and the CRS DEA estimators we find that the bias-correction procedure by Simar and Wilson (1997b) does not work well for neither the LT nor the LSW-algorithm. The empirical coverage accuracies of the bias-corrected LT and LSW intervals are lower than the coverage accuracies of the simple percentile intervals. However, the bias-correction procedure seems to be better suited for the SW intervals. The bias-corrected SW-intervals have clearly better coverage accuracy than the empirical coverage of zero(!) obtained for the simple percentile SW intervals for all sample sizes and all nominal levels considered in the study.

The results for the VRS estimator in *Table 5.1* indicate that the LSW-algorithm has the best performance. For all considered sample sizes, the LSW percentile

[9] The results are obtained using Microsoft FORTRAN compiler PowerStation, version 4.0. The LP-programs are solved using the MSIMSL revised simplex routine DDLPRS.

[10] To save space, only the average coverage accuracy of the $n$ firm-specific coverage accuracies are presented for each case. The complete set of results, containing the minimum, the maximum and the variance of the average empirical coverage accuracy, can be obtained from the author upon request.

Table 5.1: Monte Carlo simulation results for the VRS estimator

| n | LT-perc | SW-perc | LSW-perc | LT-BC | SW-BC | LSW-BC |
|---|---|---|---|---|---|---|
| Nominal level: 0.80 | | | | | | |
| 20 | 0.7068 | 0.0000 | 0.7061 | 0.6830 | 0.7281 | 0.6845 |
| 30 | 0.7680 | 0.0000 | 0.7671 | 0.7383 | 0.7334 | 0.7375 |
| 60 | 0.8280 | 0.0000 | 0.8132 | 0.7999 | 0.7148 | 0.7952 |
| Nominal level: 0.90 | | | | | | |
| 20 | 0.7879 | 0.0000 | 0.7981 | 0.7303 | 0.8237 | 0.7267 |
| 30 | 0.8439 | 0.0000 | 0.8521 | 0.7795 | 0.8540 | 0.7738 |
| 60 | 0.8979 | 0.0000 | 0.9025 | 0.8327 | 0.8744 | 0.8260 |
| Nominal level: 0.95 | | | | | | |
| 20 | 0.8361 | 0.0000 | 0.8499 | 0.7463 | 0.8397 | 0.7411 |
| 30 | 0.8896 | 0.0000 | 0.8973 | 0.7909 | 0.8736 | 0.7837 |
| 60 | 0.9324 | 0.0000 | 0.9407 | 0.8386 | 0.9032 | 0.8326 |
| Nominal level: 0.975 | | | | | | |
| 20 | 0.8456 | 0.0000 | 0.8868 | 0.7516 | 0.8446 | 0.7476 |
| 30 | 0.9018 | 0.0000 | 0.9241 | 0.7961 | 0.8790 | 0.7889 |
| 60 | 0.9542 | 0.0000 | 0.9616 | 0.8417 | 0.9088 | 0.8357 |
| Nominal level:0.99 | | | | | | |
| 20 | 0.8474 | 0.0000 | 0.9109 | 0.7528 | 0.8488 | 0.7515 |
| 30 | 0.9056 | 0.0000 | 0.9444 | 0.7979 | 0.8824 | 0.7924 |
| 60 | 0.9591 | 0.0000 | 0.9739 | 0.8431 | 0.9121 | 0.8371 |

Average coverage accuracy of the DEA bootstrap percentile (perc) and bias-corrected (BC) confidence intervals for the LT, SW and LSW algorithms.

intervals have consistently the highest coverage accuracies. For the largest sample size, $n = 60$, the intervals undercover slightly for the highest nominal levels 0.975 and 0.99 but works well for the lower levels. Furthermore, the results indicate that SW-algorithm does perform slightly better than the LT-algorithm for the smallest sample size $n = 20$. This is not surprising since the smoothing procedure involved has more effect the smaller the sample size. The LT-algorithm performs on the other hand better than the SW-algorithm for the sample sizes $n = 30$ and 60. Both LT and SW algorithms give confidence intervals with lower empirical coverage than the nominal levels, but the coverage accuracies increase with the sample size.

The results for the correctly specified CRS estimator in *Table 5.2* indicate that the LSW intervals show, as for the VRS estimator case, the best performance. For the smaller sample sizes the LSW intervals have coverage accuracy of the correct order of magnitude with a slightly *higher* coverage accuracy than the nominal levels. As the sample size increase the coverage accuracy of the LSW-intervals approaches the correct nominal level. Furthermore, the LT-algorithm performs overall better than the SW-algorithm. Even for the smallest sample size $n = 20$ the LT-intervals have approximately the correct coverage accuracy for the levels 0.80, 0.90 and 0.95. And as the sample size increase, the LT-intervals perform as well as the LSW-intervals.

As for the consistency properties of the algorithms the coverage accuracy of bootstrap confidence intervals from the LT and LSW-algorithms show a clear convergence of the empirical coverage accuracy to the nominal levels for sample sizes larger than 120. For the SW-algorithm, on the other hand, the results cast doubts on the consistency properties of the algorithm. The coverage accuracy of the SW-intervals tend to decrease with increased sample size for the lowest nominal level 0.80. The intervals slightly undercover for the smallest sample size $n = 20$ (with coverage accuracy of 0.76), but as the sample size increase, the coverage accuracy is continuously decreasing to reach a lowest level of 0.09 for the largest sample size $n = 500$. For the higher nominal levels, the coverage accuracies appear to have a hump-shaped pattern that reaches a maximum for larger sample sizes the higher the nominal level after which a considerable decrease in coverage accuracy is observed. Specifically, for the nominal level 0.90 the coverage accuracy reaches its maximum coverage (the correct level 0.90) for a sample size of $n = 60$ and then decrease steadily with increased sample size. Similar patterns are found

Table 5.2: Monte Carlo simulation results for the CRS estimator

| n | LT-perc | SW-perc | LSW-perc | LT-BC | SW-BC | LSW-BC |
|---|---|---|---|---|---|---|
| Nominal level: 0.80 | | | | | | |
| 20 | 0.8266 | 0.0000 | 0.8338 | 0.8136 | 0.7580 | 0.8331 |
| 30 | 0.8271 | 0.0000 | 0.8306 | 0.8196 | 0.7150 | 0.8403 |
| 60 | 0.8172 | 0.0000 | 0.8186 | 0.8310 | 0.6170 | 0.8454 |
| 120 | 0.8083 | 0.0000 | 0.8102 | 0.8401 | 0.4000 | 0.8509 |
| 250 | 0.8042 | 0.0000 | 0.8053 | 0.8468 | 0.2450 | 0.8535 |
| 500 | 0.8015 | 0.0000 | 0.8035 | 0.8509 | 0.0910 | 0.8552 |
| Nominal level: 0.90 | | | | | | |
| 20 | 0.9079 | 0.0000 | 0.9347 | 0.8681 | 0.8530 | 0.8798 |
| 30 | 0.9219 | 0.0000 | 0.9306 | 0.8805 | 0.8860 | 0.8845 |
| 60 | 0.9165 | 0.0000 | 0.9187 | 0.8801 | 0.9000 | 0.8825 |
| 120 | 0.9093 | 0.0000 | 0.9104 | 0.8798 | 0.8330 | 0.8820 |
| 250 | 0.9045 | 0.0000 | 0.9053 | 0.8795 | 0.5010 | 0.8813 |
| 500 | 0.9016 | 0.0000 | 0.9030 | 0.9793 | 0.2260 | 0.8809 |
| Nominal level: 0.95 | | | | | | |
| 20 | 0.9414 | 0.0000 | 0.9690 | 0.8781 | 0.8600 | 0.8825 |
| 30 | 0.9582 | 0.0000 | 0.9731 | 0.8881 | 0.8920 | 0.8913 |
| 60 | 0.9615 | 0.0000 | 0.9676 | 0.8907 | 0.9140 | 0.8912 |
| 120 | 0.9577 | 0.0000 | 0.9581 | 0.8888 | 0.9470 | 0.8900 |
| 250 | 0.9544 | 0.0000 | 0.9546 | 0.8875 | 0.8900 | 0.8894 |
| 500 | 0.9619 | 0.0000 | 0.9524 | 0.8878 | 0.4810 | 0.8892 |
| Nominal level: 0.975 | | | | | | |
| 20 | 0.9425 | 0.0000 | 0.9941 | 0.8783 | 0.8650 | 0.8831 |
| 30 | 0.9630 | 0.0000 | 0.9919 | 0.8886 | 0.8940 | 0.8914 |
| 60 | 0.9802 | 0.0000 | 0.9890 | 0.8921 | 0.9140 | 0.8941 |
| 120 | 0.9815 | 0.0000 | 0.9845 | 0.8932 | 0.9520 | 0.8943 |
| 250 | 0.9801 | 0.0000 | 0.9802 | 0.8918 | 0.9680 | 0.8933 |
| 500 | 0.9778 | 0.0000 | 0.9778 | 0.8921 | 0.8360 | 0.8934 |
| Nominal level: 0.99 | | | | | | |
| 20 | 0.9425 | 0.0000 | 0.9992 | 0.8783 | 0.8690 | 0.8833 |
| 30 | 0.9630 | 0.0000 | 0.9996 | 0.8886 | 0.9000 | 0.8914 |
| 60 | 0.9823 | 0.0000 | 0.9982 | 0.8921 | 0.9190 | 0.8942 |
| 120 | 0.9905 | 0.0000 | 0.9963 | 0.8938 | 0.9530 | 0.8956 |
| 250 | 0.9921 | 0.0000 | 0.9936 | 0.8939 | 0.9720 | 0.8955 |
| 500 | 0.9912 | 0.0000 | 0.9914 | 0.8943 | 0.9740 | 0.8956 |

Average coverage accuracy of the DEA bootstrap percentile (perc) and bias-corrected (BC) confidence intervals for the LT, SW and LSW algorithms.

for the nominal level 0.95 intervals which reaches a maximum coverage of the correct level 0.90 for a sample size of $n = 120$ and for the nominal level 0.975 intervals which has a maximum coverage of about the correct level (0.97) for a sample size of $n = 250$.

To say the least, these results cast serious doubts on the consistency properties of the SW-algorithm for computing bootstrap confidence intervals.

## 6. Summary and Conclusions

This paper performs a comparative Monte Carlo simulation study to evaluate the performance of three alternative to bootstrap the Data Envelopment Analysis (DEA) estimator of technical efficiency. The algorithms proposed by Löthgren and Tambour (1997) (LT) and Simar and Wilson (1997b) (SW) are presented and their respective performance are compared. The performance of a third bootstrap algorithm (the LSW-algorithm), based on a combination of the LT and SW algorithms, is also considered in the study. A comparison of the relative performance of the LT and LSW algorithms indicates whether any differences in performance is due to the use of a smoothed (SW) or non-smoothed (LT) resampling of pseudo-efficiencies in the bootstrap algorithm. Furthermore, a comparison of the relative performance of the SW and LSW algorithms indicates whether any differences in performance is due to the use of original data (SW) or pseudo-data (LT) in the evaluation of the bootstrap efficiency estimates.

The Monte Carlo study considers estimation of input-oriented technical inefficiency using a CRS technology as the true data generating process. Technical inefficiency is estimated using both the VRS and the CRS restricted DEA estimator. The coverage accuracy performance of simple percentile and bias-corrected bootstrap confidence intervals are considered for each of the LT, the SW and the LSW algorithms.

The results indicate that the best small-sample performance for both the VRS and the CRS DEA estimator is given by the combined LSW-algorithm. For the VRS estimator the simple LT-algorithm performs better than the more computationally demanding SW-algorithm and for the CRS estimator the LT-algorithm performs clearly better than the SW-algorithm. For the computationally more simple CRS estimator we consider larger sample sizes to assess some consistency properties of the algorithms. The results indicate that bootstrap confidence intervals from the LT and LSW algorithms show a clear convergence of the empirical coverage accuracy to the nominal levels for sample sizes larger than 120. For the

SW-algorithm, on the other hand, the results cast doubts on the consistency properties of the algorithm. The coverage accuracy for the SW-intervals decrease with increased sample size for the lower nominal level 0.80. For the higher nominal levels, the coverage accuracies appear to have a hump-shaped pattern that reaches a maximum for larger sample sizes the higher the nominal level after which a considerable decrease in coverage accuracy is observed for increased sample sizes. These findings contradict the statement by Simar and Wilson (1997b) that the proposed SW-algorithm provides a consistent bootstrap algorithm. The main argument in Simar and Wilson (1997b) relies on the use of the smoothed resampling. One might argue that the automatic bandwidth selection rule used in the Monte Carlo study is not "optimal" for the SW-algorithm, and that there are other selection rules that will render a consistent algorithm. This possible sensitivity of the SW bootstrap algorithm regarding consistency related to the choice of bandwidth selection rule is not discussed in Simar and Wilson (1997b) (or any other work by L. Simar and P. Wilson known to the author) but the obtained results from this Monte Carlo study indicate the need for this type of research. Nevertheless, as noted above, the obtained results are in favor of consistency of the LSW algorithm, despite the fact that it is based on exactly the same bandwidth selection rule as the SW-algorithm.

The main conclusion based on the conducted Monte Carlo study regarding the appropriate approach to bootstrap DEA estimators can be simply summarized as: Use a bootstrap algorithm where the bootstrap efficiency estimators are based on the pseudo-data as proposed by Löthgren and Tambour (1997), instead of the original data as in the algorithm by Simar and Wilson (1997b). Given this pseudo-data bootstrap efficiency estimator, the question of whether to used a simple or a smoothed resampling to obtain a consistent bootstrap procedure is of minor importance although the smoothed resampling has somewhat better small-sample properties.

Several extensions for future research can be identified. This study considers only the two most basic methods to construct bootstrap confidence intervals. Alternative more refined interval methods such as the median bias-corrected and the bias-corrected and accelerated intervals, as presented by Efron and Tibshirani (1993), can also be considered. Furthermore, as discussed above, the smoothing procedure applied in this paper in the SW-algorithm relies on an automatic selection of the bandwidth parameter that may not be the most efficient. Alternative selection rules such as cross-validation procedures, discussed and implemented by Simar and Wilson (1997a) in the DEA bootstrapping context, can also be considered. Yet another important extension is to consider the performance of competing bootstrap algorithms for multiple-input, multiple-output technologies since one of the primary motivations for using DEA is the ease with which multiple-input, multiple-output efficiency and productivity can be estimated.

These extensions are, however, left to future research. This study is the first study that considers an evaluation of the validity of the statistical inference obtained from competing DEA bootstrap algorithms. The mentioned extensions are important in their own respect, but the results from the simple setup used in this paper can nevertheless provide important information. The performance of an algorithm that fails in the simple setting (i.e., the SW-algorithm) should reasonably be questioned in more complicated real-world cases. On the other hand, for an algorithm that performs well in a simple setting (i.e., the LT and LSW algorithms) further evaluations are needed to assess more general robustness and validity properties.

## References


Banker, R. D. (1993). Maximum Likelihood, Consistency and Data Envelopment Analysis: A Statistical Foundation, *Management Science* 39, 1265–1273.

Charnes, A., Cooper, W. W. and E. Rhodes. (1978). Measuring the Efficiency of Decision Making Units, *European Journal of Operational Research* 2, 429–444.

Efron, B. (1979). Bootstrap methods: Another Look at the Jackknife, *Annals of Statistics* 7, 1–26.

Efron, B. and R. J. Tibshirani. (1993). *An Introduction to the Bootstrap*, Monographs on Statistics and Applied Probability 57, Chapman and Hall, London.

Farrell, M. J. (1957). The measurement of Productive Efficiency, *Journal of the Royal Statistical Society, Series A* 120, 253–281.

Ferrier, G. D. and J. G. Hirshberg (1997). Bootstrap Confidence Intervals for Linear Programming Efficiency Scores: With an Illustration Using Italian Banking Data, *Journal of Productivity Analysis* 8, 19–33.

Gstach, D. (1995). Comparing Structural Efficiency of Unbalanced Subsamples: A Resampling Adaptation of Data Envelopment Analysis, *Empirical Economics* 20, 531–542.

Kneip, A., Park, B. U. and L. Simar. (1996). A Note on the Convergence of Nonparametric DEA Efficiency Estimates, CORE, Discussion Paper 9603, Center for Operations Research and Econometrics, Université Catholique de Louvain.

Korostelev, A. P., Simar, L. and A. B. Tsybakov. (1995a). Efficient Estimation of Monotone Boundaries, *The Annals of Statistics* 23, 476–489.

Korostelev, A. P., Simar, L. and A. B. Tsybakov. (1995b). On Estimation of Monotone and Convex Boundaries, *Publications de l'Institut de l'Université de Paris* 39, 3–18.

Löthgren, M. and M. Tambour. (1997). Bootstrapping the Data Envelopment Analysis Malmquist Productivity Index, *Applied Economics*, forthcoming.

Seiford, L. M. (1996). Data envelopment analysis: The Evolution of the State of the Art (1978-1995), *Journal of Productivity Analysis* 7, 99–137.

Shephard, R. W. (1953). *Cost and Production Functions*, *Lecture Notes in Economics and Mathematical Systems* 194, Springer-Verlag. Originally published by Princeton University Press 1953.

Shephard, R. W. (1970). *The Theory of Cost and Production Functions*, Princeton University Press, Princeton.

Silverman, B. W. (1986). *Density Estimation for Statistics and Data Analyis*, Chapman and Hall Ltd, London.

Simar, L. and P. W. Wilson. (1996). Estimating and Bootstrapping Malmquist Indices, CORE, Discussion Paper 9660, Center for Operations Research and Econometrics, Université Catholique de Louvain.

Simar, L. and P. W. Wilson. (1997a). Nonparametric Tests of Returns to Scale, Mimeo, October 1997.

Simar, L. and P. W. Wilson. (1997b). Sensitivity Analysis of Efficiency Scores: How to Bootstrap in Nonparametric Frontier Models, *Management Science*, forthcoming.

Simar, L. and P. W. Wilson. (1997c). Some Problems with the Ferrier/Hirshberg Bootstrap Idea, *Journal of Productivity Analysis*, forthcoming.