INSTITUTE OF FOOD AND RESOURCE ECONOMICS
UNIVERSITY OF COPENHAGEN

MSAP Working Paper Series

No. 03/2011

Does the distribution of efficiency scores depend on the input mix?

Mette Asmild

Institute of Food and Resource Economics

University of Copenhagen

Jens Leth Hougaard

Institute of Food and Resource Economics

University of Copenhagen

Dorte Kronborg

Department of Finance

Copenhagen Business School

Does the distribution of efficiency scores depend on the input mix?

Mette Asmild
Warwick Business School
University of Warwick

Jens Leth Hougaard
Institute of Food and Resource Economics
University of Copenhagen
and

Dorte Kronborg
Center for Statistics, Department of Finance
Copenhagen Business School

Abstract

In this paper we examine the possibility of using the standard Kruskal-Wallis rank test in order to evaluate whether the distribution of efficiency scores resulting from Data Envelopment Analysis (DEA) is independent of the input (or output) mix.

Recently, a general data generating process (DGP) suiting the DEA methodology has been formulated and some asymptotic properties of the DEA estimators have been established. In line with this generally accepted DGP, we formulate a conditional test for the assumption of mix independence. Since the DEA frontier is estimated, many standardl assumptions for evaluating the test statistic are violated. Therefore, we propose to explore its statistical properties by the use of simulation studies. The simulations are performed conditional on the observed input mixes. The method, as shown here, is applicable for models with multiple inputs and one output with constant returns to scale when comparing distributions of efficiency scores in two or more groups.

The approach is illustrated in an empirical case of demolition projects where we reject the assumption of mix independence. This means that it is not

meaningful to perform a complete ranking of the projects based on their efficiency score. Thus the example illustrates how common practice can be inappropriate.

Keywords: Data Envelopment Analysis (DEA), homogeneous efficiencies, small sample properties, Kruskal-Wallis, ranking, demolition projects

Correspondence: Dorte Kronborg, Center for Statistics, Copenhagen Business School, Solbjerg Plads 3, 2000, Frederiksberg, Denmark. E-mail: dk.mes@cbs.dk

1 Introduction

Many efficiency studies employing the nonparametric Data Envelopment Analysis (DEA) technique use the resulting efficiency scores to make a complete ranking of the set of observed units (see e.g. Adler et al., 2002 for a review). Yet, it is questionable whether such a comparison of efficiency scores is appropriate, c.f. e.g. the discussion in Bogetoft and Otto (2011). By comparing units with benchmarks on different facets of the efficient frontier these units are in effect compared based on an evaluation using different underlying (shadow) prices. The theoretical requirement for making a complete ranking is clear: comparisons of efficiency scores only make sense if the distribution of efficiency scores is independent of the input and output mix; a requirement we here call the assumption of *mix independence*. It is, however, less obvious how to test this assumption empirically. In the present paper we suggest a statistical test for whether the hypothesis of mix independence can be rejected.

From a theoretical point of view the efficiency scores have a straightforward interpretation for each observation taken separately, namely the factor by which we can scale input (costs), or outputs (revenue), in order to become as efficient as those observations spanning the frontier (the common benchmark). However, it is not obvious that efficiency scores can be compared directly between observations for which there is no dominance relation. Such units will typically be benchmarked against different facets of the frontier (or dually against different underlying weights (shadow prices)) and, strictly speaking, such a comparison is economically meaningless.

Thus, any comparison of efficiency scores for observations with different input (or output) mixes does, in fact, rely on the hypothesis of mix independence, i.e. that the distribution of efficiencies is independent of input (or output) mix. If the mix independence assumption is violated, we know that there are differences between the efficiency distributions for different input mixes, which means that it is not appropriate to compare efficiency scores across the sample or, for instance, rank all observations based on their efficiency scores.

When using DEA, an estimate of the true but unknown production function is obtained from a convex envelopment of the observed data points (Charnes et al., 1978). Banker (1993) was the first to show that the estimated DEA frontier is a consistent maximum likelihood estimator within a certain class of sensible functions. Since then, statistical properties of the efficiency estimators have been subject to numerous studies. Simar and Wilson (2000a) summarize the most important recent results, among which results on consistency and convergence rates for DEA estimators are prominent.

It is well known that both the estimate of the production frontier and the efficiency scores are downward biased. The distribution of the estimated efficiency scores

is unknown except from the very special case of one input and one output. This case is, however, too simple for practical applications and much effort has been put into development of consistent bootstrap methods for both the estimated DEA frontier and the estimated efficiencies in the general case (Simar and Wilson, 2000a, 2000b). Kneip et al. (2008) find an expression for the asymptotic distribution of the empirical inefficiencies in the general variable returns to scale (VRS) set-up. However, the distribution depends on several unknown values, and they suggest consistent bootstrap methods to obtain quantiles for the 'unknown' distribution.

Recently, Simar and Zelenyuk (2006) investigated the possibility of using the test of Li (1996, 1999) for similarity of two unknown distributions within a DEA context. They show that a modification of Li's test can be useful for comparing distributions of efficiencies based on the estimated scores. However, Li's test is restricted to a comparison of two densities. Therefore, there is a need for investigating the possibility of using tests for comparison of more than two density functions.

The aim of this paper is to investigate the properties of the popular Kruskal-Wallis test for comparison of distributions of efficiency scores. Specifically we here consider a multiple-input one-output DEA model with constant returns to scale and investigate how the uncertainty arising from estimation of the frontier and the efficiency scores influence the statistical properties of the test statistic. The proposed method is illustrated using an empirical dataset of 169 demolition projects.

In section 2 below, the theoretical model is introduced followed by a description of the dataset considered. In section 4 the Kruskal-Wallis test statistic is presented together with the simulation approach designed to investigate its properties. Simulation results are presented in section 5, followed by a discussion of the applicability of the proposed method as well as some further practical considerations.

2 The Model

Consider production plans (X, Y) where r inputs, $X \in \mathbb{R}^r_+$, are used to produce one output, Y, such that $(X, Y) \in \mathbb{R}^{r+1}_+$. Let $Z \in \mathbb{R}^r_+$ describe the output scaled production plan (i.e. $Z = X/Y$). Further, let $V = \|Z\|$ be the Euclidian norm of Z and $U = \frac{Z}{\|Z\|}$ be the direction of Z (U is on the unit sphere in $\mathbb{R}^r_+$) such that

$$Z = U \cdot V.$$

Let $g(\cdot) : \mathbb{R}^r_+ \to \mathbb{R}$ be a homogeneous and convex function such that the input set becomes,

$$L = \{z \in \mathbb{R}^r_+ \mid g(z) \geq 1\},$$

and g represents the isoquant (or production frontier). Let Shephard's distance function (Shephard, 1970) E_Z of Z relative to the isoquant g be defined as,

$$E_Z = \sup\{e \in \mathbb{R}_+ \mid g(\frac{Z}{e}) \geq 1\} = g(Z).$$

Since,

$$E_Z = g(U \cdot V) = g(U) \cdot V,$$

Z can be decomposed into

$$Z = U \cdot V = U \cdot \frac{E_Z}{g(U)}. \tag{1}$$

Note that $E_Z \geq 1$ and that $E_Z^{-1} \in [0, 1]$ corresponds to Farrell's index of technical input efficiency, see e.g., Farrell (1957). An illustration of the concepts is provided in section 4 below.

We consider a data generating process, DGP, where firms first choose an input mix (or direction) U and then, given U, choose an efficiency score E_Z, in line with Simar and Wilson (1998, 2000a). Given this GDP, the joint density $f_{(E_Z,U)}(\cdot,\cdot)$, is naturally decomposed into the conditional density given U and the marginal density for U:

$$f_{(E_Z,U)}(e,u) = f_{(E_Z|U)}(e|u) f_U(u).$$

It is a fundamental (though often implicit) assumption when comparing efficiency scores between firms that the density distribution of E_Z is independent of the input mix U,

Mix Independence: $\quad f_{(E_Z|U)}(e|u) = f_{E_Z}(e).$

In the following we suggest a statistical method for testing this assumption namely independence between E_Z and U, $E_Z \perp\!\!\!\perp U$. We consider a method building on the proposed DGP for Z on L.

3 Empirical illustration

To illustrate our approach we consider a data set from a large demolition company. The data set consists of 169 different demolition projects undertaken by the company within one year. Each project uses combinations of labor costs, machine costs and other variable costs, which constitute the inputs in the efficiency assessment. The single output considered is the revenue generated by the projects. While this

represents a cost function rather than a production function as such, the interpretation of the efficiencies is similar to those from the standard production model above. Projects with a zero cost on any of the inputs have been excluded. Note also that the total costs in some cases are higher than the revenue. Descriptive statistics of the variables are given in Table 1 below.

	Mean	Std.dev	Min	Max
Revenue	1737710	2509220	27396	13247406
Machine costs	241483	484542	250	3559392
Labor costs	330887	741288	235	6054610
Other costs	907937	1359713	2229	9469283
Efficiency	0.563	0.200	0.161	1

Table 1. Descriptive statistics for output and input variables as well as for efficiency scores.

To assess the efficiencies of the individual projects, an input-oriented DEA model with constant returns to scale (the CCR model) was used to estimate the production function and corresponding input efficiency scores (c.f. Charnes et al., 1978). The descriptive statistics of these efficiency scores are provided in the last row in Table 1 above.

Often analysts as well as practitioners are interested in ranking the observed units on the basis of their efficiency scores. A related issue is whether it is appropriate to directly compare two efficiency scores from different parts of the production space. When analyzing the demolition projects the question arose whether labor intensive projects are more or less efficient than machine intensive projects. Answering this question requires a comparison of efficiency scores between projects with very different characteristics, for instance removing asbestos panels in a working hospital, which is very labor intensive, and the complete demolition of a multi story car park, which is very machine intensive. Therefore we here propose an approach for testing whether such comparisons are, in fact, appropriate.

For the actual test for mix independence described in detail in Section 5 below, the data set must be partitioned into a number of distinct subgroups based on their input mixes. In general, the partitioning into groups has to be economically meaningful in the sense that directions corresponds to production activities and defining relevant groups for comparison depends on the characteristics of the specific data set at hand. An obvious partitioning in the current case would come from considering the

largest cost component for each project, where the groups then directly indicate whether the projects are mainly labor-, machine- or 'other cost'-intensive. In the current data set this would, however, result in most of the projects being categorized as 'other cost' intensive, since that variable is generally on a larger scale than the other two input variables, c.f. also Table 1. In the following we have considered two different partitionings of the current data set into three disjoint cones in the production space. In both partitionings the groups reflect labor-, machine and other-cost intensive projects respectively and with a reasonable number of projects in each group:

Partitioning A: The output scaled input variables Z are each divided by the maximum value on the given variable such that they are in the same order of magnitude. For each observation, the group is determined by the largest of these transformed variables values, resulting in 1) 103 observations with largest values on other costs, 2) 40 observations with largest machine costs and 3) 26 observations with largest labor costs.

Partitioning B: To ensure a more even number of observations in all groups than above, the second partitioning is: 1) Other costs/revenue above the 66.7% quantile (56 projects), 2) If not in group 1 and machine cost/revenue larger than salary/revenue (46 projects), and 3) If not in group 1 and machine cost/revenue smaller than salary/revenue (67 projects).

4 Testing mix independence

Assume that the frontier g is known and denote by z_i n independent observations of Z from $i = 1, \ldots, n$ production plans. Let $((U_1, E_1), \cdots, (U_n, E_n))$ be stochastic variables denoting directions and efficiencies for the n production plans, and (u_i, e_i), $i = 1, \ldots, n$ be the corresponding observed directions and efficiencies. According to (1), z_i can be decomposed as $z_i = u_i v_i = u_i e_i g(u_i)^{-1}$ for $i = 1, \ldots, n$. The hypothesis $E \perp\!\!\!\perp U$, can be evaluated by use of a conditional test. Given mix independence and given the directions, the (in)efficiency scores are identically distributed. The proposed test statistic is based on the observed directions u_i and efficiencies e_i. A (conditional) test for mix independence based on the (in)efficiencies can be calculated as an ordinary non-parametric rank test comparing distributions between different direction-based cones of the input set, see Figure 1 below.

The figure shows, for a given point Z_i, the corresponding direction represented by the point U_i on the unit sphere and the length V_i. It also illustrates a partitioning of the production space in three disjoint cones (I,II,III) based on the input mix.

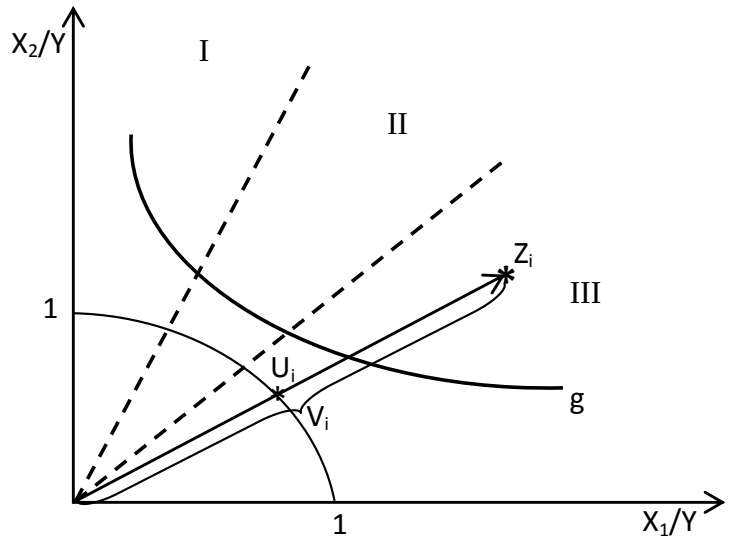

Figure 1. Illustration of the notation and various elements.

4.1 Kruskal - Wallis test

Let the input set L be partitioned into k cones corresponding to a partitioning of the units sphere in $\mathbb{R}^r_+$. Let n_j be the number of observations in the j'th cone $j = 1, \ldots, k$ $(n_1 + \cdots + n_k = n)$ and let the (in)efficiency measures within that cone be numbered $e_{n_1+\cdots+n_{j-1}+1}, \ldots, e_{n_1+\cdots+n_{j-1}+n_j}$ and have density f_j. For notational convenience, the set of indices $\{n_1+\cdots+n_{j-1}+1, \ldots, n_1+\cdots+n_{j-1}+n_j\}$ is denoted s_j.

Let $R_1, \ldots, R_n$ be the ranks based on e_i in the total sample. Ranks are well-defined as long as the probability of coincidence is zero. Assuming that the ranks are well-defined the k-sample Kruskal-Wallis (KW) test for comparison of k densities (Hájek and Šidák (1965), Lehmann (1974)) is

$$Q = \frac{12}{n(n+1)} \sum_{j=1}^{k} \frac{1}{n_j} (\sum_{l \in s_j} R_l)^2 - 3(n+1).$$

Under the assumption of equal densities (H_0: $f_j = f,\ \forall j$) the KW test statistic is asymptotically χ^2-distributed with $k-1$ degrees of freedom when $\min(n_1, \ldots, n_k) \to \infty$. Ties are (as usual) handled by assigning the average rank to a group of tied values. A correction for ties can be made by dividing Q by $1 - \frac{\sum_{m=1}^{h}(\tau_m^3 - \tau_m)}{n^3 - n}$, where h is the number of groupings of different tied ranks, and $\tau_1, \ldots, \tau_h$ are sizes of ties.

Unless there are a large number of ties the correction usually only makes a little difference.

Calculating the KW test statistic for the two partitionings suggested in Section 4 above gives the results shown in Table 2 below. The immediate interpretation of the results in Table 2 implies that there is not mix independence amongst the demolition projects, i.e., it is not appropriate to rank the projects based on their efficiency scores, nor is it possible to directly compare any two efficiency scores (from projects with different input mixes).

Partitioning	KW	Df	p-value
A	8.185	2	0.0167
B	23.66	2	≤ 0.0001

Table 2. KW test statistics and corresponding p-values originating from the $\chi^2(2)$-distribution.

However, in practice the production function g is unknown and when using an estimated production frontier the assumptions required for the asymptotic properties of the KW test statistics are violated. Therefore, the p-values shown above are not necessarily reliable. By using simulation studies we aim to investigate the impact on the distribution of the KW test statistic of using the DEA estimated production technology $\hat{g}$.

4.2 Simulation of the distribution of the test statistic

1. Let $(u_1, \ldots, u_n)$ be the observed directions. Let a_i be the number of observations in direction i (typically a_i will be equal to 1). Fix this set of observed directions.

2. Assume a "true" technology g, for example by specifying parameters in a Cobb-Douglas functional form.

3. Given the assumed "true" technology g, for each observed direction u_i simulate a_i data points $\tilde{z}_i$ by drawing an inefficiency score $\tilde{e}_i^{-1}$ from a suitable distribution f on $[0, 1]$, i.e. $\tilde{z}_i = u_i \tilde{e}_i g(u_i)^{-1}$.

4. Use the simulated data points $\tilde{z}_i$ to make a CCR estimate $\hat{g}$ of the production function and determine the efficiency score $\hat{e}_i$ for each point relative to $\hat{g}$.

5. Based on a partitioning of the input space into k cones calculate the associated Kruskal-Wallis test statistic.

6. For fixed f and k repeat this procedure (step 3, 4 and 5) a number of times (N=10000 say).

In order to get a consistent estimate of the production frontier, g, f has to be chosen such that there is a positive probability for observing production plans arbitrarily close to the boundary as n gets large, c.f. e.g. Kneip et al. (1998). That is, f has to satisfy the following condition:

For all directions u there exists constants $\varepsilon_1 > 0$ and $\varepsilon_2 > 0$ such that for all $e \in [g(u), g(u) + \varepsilon_2]$ we have $f(e|u) \geq \varepsilon_1$.

5 Simulation results

We have chosen to consider four different shapes of Cobb-Douglas production functions with constant returns to scale ($z_1^{\alpha_1} z_2^{\alpha_2} z_3^{\alpha_3}$, $\alpha_1 + \alpha_2 + \alpha_3 = 1$) where z_1 is the revenue scaled machine costs, z_2 is the revenue scaled labor costs and z_3 the revenue scaled other costs. The α parameters considered are $(\alpha_1, \alpha_2, \alpha_3) = [(\frac{1}{3}, \frac{1}{3}, \frac{1}{3}), (\frac{1}{10}, \frac{1}{10}, \frac{8}{10}), (\frac{8}{10}, \frac{1}{10}, \frac{1}{10}), (0.2, 0.15, 0.65)]$ where the latter is chosen near the ordinary least square (OLS) estimated parameters restricted to summing to one. For each Cobb-Douglas specification we consider different efficiency distributions all with relatively high probability mass in the neighborhood of 1, specifically we use various Beta distributions, Beta(1,1), Beta(3,1), Beta(2, 0.8) and Beta(5, 1.5), and one truncated normal distribution (TNF) $N(1, (0.2)^2)$.

For each combination of Cobb-Douglas parameters, efficiency distribution and partitioning we have performed 10000 simulations of efficiency scores with one simulated observation in each observed direction, the results of which are shown in Table 3 below. The first column shows the significance probabilities in the actual χ^2-distribution. Each value in the interior of the table represents, for the significance probability in the given row, the corresponding significance probability from the simulated distribution. The rows named "Observed" give the significance probabilities based on the simulations for the obtained KW test statistics.

Considering first the bottom section of Table 3 where the OLS estimated Cobb-Douglas parameters are used, we notice that the simulated upper tail distributions for the KW test statistic are generally very similar to those from the χ^2-distribution. This is the case for all the efficiency distributions considered, but for partitioning B the results are generally closer to the χ^2-distribution than those from partitioning A. This is likely due to the fact that partitioning A has a fairly uneven split of observations between the groups, and the parts of the estimated frontier determined by the observations in the smaller groups will tend to vary more, resulting in less

$\chi^2(2)$ significance probability	Partitioning A $(\alpha_1,\alpha_2,\alpha_3)=(\frac{1}{3},\frac{1}{3},\frac{1}{3})$ Beta(λ_1,λ_2) (1,1)	(3,1)	(2,0.8)	(5.1.5)	TNF (1,$(0.2)^2$)	Partitioning B $(\alpha_1,\alpha_2,\alpha_3)=(\frac{1}{3},\frac{1}{3},\frac{1}{3})$ Beta(λ_1,λ_2) (1,1)	(3,1)	(2,0.8)	(5.1.5)	TNF (1,$(0.2)^2$)
0.100	.0956	.1391	.1154	.1695	.1872	.0687	.0778	.0722	.0937	.0962
0.050	.0438	.0709	.0551	.0855	.0955	.0301	.0342	.0308	.0417	.0415
0.025	.0211	.0367	.0261	.0423	.0478	.0140	.0138	.0129	.0188	.0167
0.010	.0070	.0137	.0095	.0147	.0178	.0057	.0045	.0037	.0060	.0057
0.001	.0006	.0012	.0003	.0016	.0011	.0002	.0001	.0004	.0002	.0001
Observed	.0145	.0243	.0163	.0272	.0306	0	0	0	0	0
$\chi^2(2)$ significance probability	$(\alpha_1,\alpha_2,\alpha_3)=(\frac{1}{10},\frac{1}{10},\frac{8}{10})$ Beta(λ_1,λ_2) (1,1)	(3,1)	(2,0.8)	(5.1.5)	TNF (1,$(0.2)^2$)	$(\alpha_1,\alpha_2,\alpha_3)=(\frac{1}{10},\frac{1}{10},\frac{8}{10})$ Beta(λ_1,λ_2) (1,1)	(3,1)	(2,0.8)	(5.1.5)	TNF (1,$(0.2)^2$)
0.100	.0725	.0646	.0688	.0715	.0837	.1339	.1377	.1231	.1492	.1343
0.050	.0327	.0279	.0280	.0297	.0406	.0666	.0691	.0590	.0780	.0678
0.025	.0143	.0121	.0129	.0123	.0166	.0323	.0368	.0284	.0381	.0328
0.010	.0049	.0037	.0037	.0033	.0047	.0128	.0142	.0099	.0154	.0107
0.001	.0003	.0004	.0000	.0002	.0002	.0011	.0013	.0009	.0007	.0008
Observed	.0090	.0076	.0080	.0075	.0099	0	0	0	0	0
$\chi^2(2)$ significance probability	$(\alpha_1,\alpha_2,\alpha_3)=(\frac{8}{10},\frac{1}{10},\frac{1}{10})$ Beta(λ_1,λ_2) (1,1)	(3,1)	(2,0.8)	(5.1.5)	TNF (1,$(0.2)^2$)	$(\alpha_1,\alpha_2,\alpha_3)=(\frac{8}{10},\frac{1}{10},\frac{1}{10})$ Beta(λ_1,λ_2) (1,1)	(3,1)	(2,0.8)	(5.1.5)	TNF (1,$(0.2)^2$)
0.100	.1206	.1410	.1191	.1742	.1785	.0910	.0763	.0773	.0862	.0787
0.050	.0595	.0741	.0626	.0908	.0959	.0441	.0333	.0337	.0387	.0353
0.025	.0300	.0368	.0289	.0462	.0472	.0205	.0151	.0143	.0184	.0159
0.010	.0117	.0151	.0109	.0217	.0183	.0066	.0041	.0049	.0064	.0051
0.001	.0012	.0008	.0008	.0021	.0018	.0005	.0003	.0008	.0006	.0002
Observed	.0196	0252	.0188	.0325	.0302	0	0	0	0	0
$\chi^2(2)$ significance probability	$(\alpha_1,\alpha_2,\alpha_3)=(0.20,0.15,0.65)$ Beta(λ_1,λ_2) (1,1)	(3,1)	(2,0.8)	(5.1.5)	TNF (1,$(0.2)^2$)	$(\alpha_1,\alpha_2,\alpha_3)=(0.20,0.15,0.65)$ Beta(λ_1,λ_2) (1,1)	(3,1)	(2,0.8)	(5.1.5)	TNF (1,$(0.2)^2$)
0.100	.0744	.0819	.0781	.0958	.1138	.1094	.1110	.1048	.1324	.1228
0.050	.0321	.0362	.0329	.0455	.0551	.0592	.0500	.0500	.0592	.0602
0.025	.0145	.0139	.0140	.0198	.0251	.0305	.0247	.0238	.0305	.0302
0.010	.0039	.0038	.0045	.0039	.0078	.0118	.0103	.0071	.0118	.0111
0.001	.0001	.0003	.0001	.0001	.0003	.0008	.0005	.0003	.0008	.0009
Observed	.0089	.0077	.0075	.0089	.0157	0	0	0	0	0

Table 3. Simulation results based on 10000 simulations.

stable ranks. In partitioning B, however, the observations are much more evenly split between the groups such that there are more observations in the smallest group. Furthermore it appears that the more the shape of the frontier differs from that de-

termined by the actual data points, the larger the simulated probabilities deviate from those from the χ^2-distribution. However, no matter what, the probability distributions are still very similar, and with the results from partitioning B generally slightly closer to the χ^2-distribution than those from partitioning A. Considering finally the last row in each section of the table, we can conclude that for all combinations of partitioning, Cobb-Douglas parameters and efficiency distribution, the significance probability for the hypothesis of mix independence for the demolition data set is less than around 3 percent, which leads to the hypothesis being rejected in all cases. Additionally, in this empirical case, the simulated results are not substantially different from those from the χ^2-distribution (see Table 2) and lead to the same conclusion about the (lack of) mix independence. Therefore, the χ^2-distribution could actually have been used directly in this case, though we would not have known so without the information from the simulation study.

6 Discussion

In this paper we have proposed a method to investigate the hypothesis of mix independence, that is, whether the distributions of efficiency scores are the same for different input mixes. This is important whenever, for instance, a ranking of observations based on efficiency scores is desired. Where the test of Li (1996, 1999), adapted to the DEA context by Simar and Zelenyuk (2006), compares two pre-existing groups in the data set, our test is based on partitioning the production space into any number of non-overlapping cones defined from the input and output mixes (directions). Thus, the two tests are fundamentally different and address different research questions.

Our method is particularly relevant whenever one suspects that the efficiencies depend on the input mix, in the demolition case for example that labor intensive projects might be less efficient than machine intensive projects, or when there are no exogenously defined groupings of the data set.

The test for mix independence utilizes a standard Kruskal-Wallis test, available in most statistical software packages. The KW test is a rank test which does not rely on questionable distributional assumptions and is generally robust (Wei, 1981). The asymptotic properties of the KW test are well-known, with the test statistic being asymptotically χ^2-distributed. But it has not been investigated how the KW test statistic behaves when calculated from efficiency scores based on an empirical

estimation of the production function in place of the true but unknown function. Therefore simulation studies, like the ones presented here, should be used in practice to estimate the empirical distribution of the test statistic and corresponding significance probabilities.

In our empirical case with 169 observations, the simulation results show that the distribution of the test statistic is, in fact, reasonably close to a χ^2-distribution. Therefore, for practical purposes, the standard asymptotic results could have been used and would here have led to the same conclusion: there is not mix independence in the demolition data set. Consequently, it is not appropriate to rank the projects nor possible to determine whether labor intensive projects are less efficient than machine intensive projects or in other ways directly compare efficiency scores for projects with different input mixes (unless, of course, there is a direct dominance relation between the projects in question).

In our simulation study we utilize different partitionings of the data set, different shapes of the production function as well as different distributional assumptions for the inefficiencies often considered in the literature. The results indicate that the empirical distribution is closest to the χ^2-distribution if the observations are fairly evenly split between the cones and if the shape of the production function used resembles that of the observed data points. This will mainly be an issue for small samples. Where the most even partitioning of observations resulted in the best approximation, neither of the two partitionings shown previously yielded substantial deviations of the simulated distributions from the χ^2-distribution even in the present small sample study. However, further simulations with extremely uneven splits of observations between groups (11,18,140 respectively) resulted in an empirical distribution very far from the expected χ^2-distribution. So if the interest is in comparing groups that happen to be of very different sizes, especially in small samples where there subsequently will be very few observations in the smallest group, simulation studies are definitely required. Otherwise, we suggest to consider fairly evenly sized groups. Similarly, even if none of the shapes of the production function investigated resulted in large deviations, the best results were obtained when using the shape closest to the observed data, which is what we would recommend for future studies.

It can also be noted that in order to investigate the large-sample properties of the test statistic, we performed simulations with five observations in each observed direction. As expected, since the estimated production function for large data sets closely resembles the true production function, the results were basically identical to the χ^2-distribution, wherefore the standard KW test can be used directly. Even in moderately sized data set, conclusions can still be drawn without simulation studies, if the significance probabilities are either very small or very large.

We close with a few remarks on the possibility of generalizing our test from the multiple input-one output constant returns to scale version presented here. First, we could consider extending to multiple outputs. Second, we could extend to variable returns to scale. In case of the latter it seems that the test developed in the present paper, in theory, easily generalizes to isoquants for each output level. In practice however, most data sets will not contain enough observations at each output level to perform such a test. The obvious solution would be to combine individual output levels into groups but this procedure still requires a substantial number of observations and relies on the existence of natural cut-off points for the output groups. In case of the former it seems less obvious how to generalize the model above. This will remain a topic for further research.

Acknowledgement: The authors would like to thank Kristiaan Kerstens for valuable comments.

References

Adler, N., Friedman, L. and Sinuany-Stern, Z. (2002). Review of ranking methods in the data envelopment analysis context. *European Journal of Operational Research*, 140, 249-265.

Banker, R.D. (1993). Maximum likelihood, consistency and data envelopment analysis: A statistical foundation. *Management Science*, 39(10), 1265-1273.

Bogetoft, P. and Otto, L. (2011). **Benchmarking with DEA, SFA and R**. Springer.

Charnes, A., Cooper, W.W. and Rhodes, E. (1978). Measuring the Inefficiency of Decision Making Units. *European Journal of Operational Research* 2(6), 429-444.

Farrell, M.J. (1957). The Measurement of Produtive Efficiency. *Journal of the Royal Statistical Society*, 120, 253-281.

Hájek, J. and Šidák, Z (1965). **Theory of Rank Tests**. Academic Press.

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

Kneip, A.,Simar, L. and Wilson, P.W. (2008). Asymptotics and consistent bootstrap for DEA estimators in nonparametric frontier models. *Econom*tric Theory* 24, 1663-1697.

Lehmann, E.L. (1974). **Nonparametrics. Statistical methods based on Ranks**. Springer.

Li, Q. (1996). Nonparametric testing of closeness between two unknown distribution functions. *Econometric Reviews*, 15, 261-274.

Li, Q. (1999). Nonparametric testing of the similarity of two unknown density functions: local power and bootstrap analysis. *Nonparametric Statistics*, 11, 189-213.

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

Simar, L. and Wilson, P.W. (1998). Sensitivity Analysis of Efficiency Scores: How to Bootstrap in Nonparametric Frontier Models. *Management Science*, 44 (1), 49-61.

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

Simar, L. and Wilson, P.W. (2000b). A general methodology for bootstrapping in non-parametric frontier models. *Journal of Applied Statistics*, 13, 49-78.

Simar, L. and Zelenyuk, V. (2006). On testing equality of distributions of technical efficiency scores. *Econometric Reviews*, 25(4), 497-522.

Wei, L.J. (1981). Asymptotic Conservativeness and Efficiency of Kruskal-Wallis Test for K Dependent Samples. *Journal of the American Statistical Association*, 76, 1006-1009.