ECONOMIC STUDIES

DEPARTMENT OF ECONOMICS
SCHOOL OF ECONOMICS AND COMMERCIAL LAW
GÖTEBORG UNIVERSITY
141

FOUR ESSAYS ON THE MEASUREMENT OF PRODUCTIVE EFFICIENCY

Dag Fjeld Edvardsen

ISBN 91-85169-00-5
ISSN 1651-4289 print
ISSN 1651-4297 online

Four Essays on the Measurement of Productive Efficiency

Doctoral thesis by Dag Fjeld Edvardsen (dfe@byggforsk.no)

Preface

After graduating in economics from the University of Oslo I started working as a research assistant at the Frisch Centre in January 1999. My first task was trying to understand a strange method I had never heard of before. It was refereed to as Data Envelopment Analysis (DEA). Soon I was working with Finn R. Førsund and Sverre A.C. Kittelsen on applied projects where DEA was used to measure technical efficiency. Examples were nursing homes and home care, employment offices, colleges, electricity distribution utilities, and physical therapists.

In 2000 Norwegian Building Research Institute (NBI) in cooperation with the Frisch Centre wrote an application to the Norwegian Research Council (NFR). The topic to be investigated was the efficiency of the Norwegian construction industry. When the application was accepted I was hired at NBI as a doctoral student.

I would like to thank NFR for financing the three years it has taken to write the four essays in this thesis. I am deeply grateful to NBI for offering me the opportunity to be part of the project “Productivity in Construction” (and for providing a very large amount of coffee). Frank Henning Holm (now head of NBI) and Grethe Bergly (now at Multiconsult) deserve thanks for hiring me. In 2001 Thorbjørn Ingvaldsen became leader of the project when Grethe Bergly went to Multiconsult. His encouragement, humour, and patients have been enormous, as is his knowledge of the Norwegian construction industry. Jon Rønning became head of PROS (the department this project is located at) when Frank Henning Holm left to become head of NBI. Jon’s support has been without parallel. I would also like to thank my patient and understanding colleges at NBI for encouragement.

My thesis advisors Lennart Hjalmarsson and Finn R. Førsund have been the best advisors a doctoral student could ever wish for. Lennart has been very supportive and helped me every time things seemed very difficult. Finn has always been there for me, and his advice and support have been a necessary condition for this thesis to exist. Finn and Lennart’s knowledge of microeconomics and efficiency analysis are without doubt world class.

Sverre A.C. Kittelsen (Frisch Centre) has been an enormous resource for me. His knowledge of the subtle and difficult parts of efficiency analysis is more than impressive. He has also been invaluable when it comes to the development of the software used for the bootstrap calculations used in this thesis.

The members of the reference group for “Productivity in Construction” deserve thanks for understanding that some task are worth doing even if they require time: Rolf Albriktsen (Veidekke ASA), Finn R. Førsund (University of Oslo/Frisch Centre), Frank Henning Holm (NBI), Sverre Larsen (BNL), Knut Samset (NTNU), Arild Thommasen (Statistics Norway at Kongsvinger), and Grethe Bergly (Multiconsult).

Last, but not least, I would like to thank my family: My late mother Laila (who died last year), my father Johny, my sister Janne, my aunt Unni, and my uncle Hugo. Their support has been invaluable, and without it this thesis would not exist.

Oslo, November 2004

Dag Fjeld Edvardsen

Contents

Abstract

This collection of essays contains two kinds of contributions. All four essays include applications of the existing DEA (Data Envelopment Analysis) toolbox on real world datasets. But the main contribution is that they also offer new and useful tools for practitioners doing efficiency and productivity analysis.

Essay I is about benchmarking by means of applying the DEA model on electricity distributors. A sample of large electricity distribution utilities from Denmark, Finland, Norway, Sweden and the Netherlands for the year 1997 is studied by assuming a common production frontier for all countries. The peers supporting the benchmark frontier are from all countries. New indices describing cross-country connections at the level of individual peers and their inefficient units as well as between countries are developed, and novel applications of Malmquist productivity indices comparing units from different countries are performed.

The contribution of Essay II is to develop a method for classifying self-evaluators based on the additive DEA model into interior and exterior ones. The exterior self-evaluators are efficient “by default”; there is no firm evidence from observations for the classification. These units should therefore not been regarded as efficient, and should be removed from the observations of efficiency scores when performing a two-stage analysis of explaining the distribution of the scores. The application to municipal nursing- and home care services of Norway shows significant effects of removing exterior self-evaluators from the data when doing a two-stage analysis.

The robustness of the efficiency scores in DEA has been addressed in Essay III. It is of crucial importance for the practical use of efficiency scores. The purpose is to demonstrate the usefulness of a new way of getting an indication of the sensitivity of each of the efficiency scores to measurement error. The main idea is to investigate a DMU’s (Decision Making Unit) sensitivity to sequential removal of its most influential peer (with new peer identification as a part of each of the iterations). The Efficiency stepladder approach is shown to provide relevant and useful information when applied on a dataset of Nordic and Dutch electricity distribution utilities. Some of the empirical efficiency estimations are shown to be very sensitive to the validity and existence of one or a low number of other observations in the sample. The main competing method is Peeling, which consists of removing all the frontier units in each step. The new method has some strengths and some weaknesses in comparison. All in all, the Efficiency stepladder measure is simple and crude, but it is shown that it can provide useful information for practitioners about the robustness of the efficiency scores in DEA.

Essay IV is an attempt to perform an efficiency study of the construction industry at the micro level. In this essay information on multiple outputs is utilized by applying DEA on a cross section dataset of Norwegian construction firms. Bootstrapping is applied to select the scale specification of the model. Constant returns to scale was rejected. Furthermore, bootstrapping was used to estimate and correct for the sampling bias in the DEA efficiency scores. One important lesson that can be learned from this application is the danger of taking the efficiency scores from uncorrected DEA calculations at face value. A new contribution is to use the inverse of the standard errors (from the bias correction of the efficiency scores) as weights in a regression to explain the efficiency scores. Several of the hypotheses investigated concerning the latter are found to have statistically significant empirical relevance.

Introduction

A key paradigm in neo-classical production theory is that firms operate on the production frontier. However, even a superficial observation of real production units indicates that this is most often not the case. It is then rather odd that economists continue to believe in this paradigm, and that so little effort is spent on revealing inefficiencies and their causes. Most of the old tricks learned in microeconomics become invalid since they adopt the assumption from neo-classical economics that firms behave as if they were technically efficient.

A natural starting point for developing methods for the study of productive efficiency is the seminal 1957 paper by Michael J. Farrell with the appropriate title "The measurement of productive efficiency." Farrell's key contribution was introducing a non-parametric method for estimating the efficient production frontier as a reference for his efficiency measures, based on enveloping data "from above." This approach generalizes naturally to multiple inputs and multiple outputs.

The four essays in this thesis are modest attempts to follow up the Farrell tradition as it has been developed both within economics and operations research where the term DEA was coined in Charnes *et al.* (1978).

Essay I. International benchmarking of electricity distribution utilities[1]

Improvement of efficiency in electricity distribution utilities has come on the agenda, as an increasing number of countries moved towards deregulation of the sector in the last decade. A key element in assessing potentials for efficiency improvement is to establish benchmarks for efficient operation. A standard definition of benchmarking is a comparison of some measure of actual performance against a reference performance. One way of obtaining a comprehensive benchmarking as opposed to partial key ratios is to establish a frontier production function for utilities, and then calculate efficiency scores relative to the frontier.

In this study a piecewise linear frontier is used, and technical efficiency measures (Farrell, 1957) and Malmquist productivity measures (Caves et al., 1982) are calculated by employing the DEA model (Charnes et al., 1978). The DEA model has been used in several studies of the utilities sector recently. A special feature of the present cross section study is that the data (for 1997) is based on a sample of utilities from five countries: Denmark, Finland, The Netherlands, Norway and Sweden. Most of the efficiency studies of utilities have been focusing on utilities within a single country (Førsund and Kittelsen, 1998), but a few studies have also compared utilities from different countries (Jamasb and Pollitt, 2001). In some cases an international basis for benchmarking is a necessity due to the limited number of similar firms within a country. When the number of units is not the key motivation for an international sample for benchmarking, the motivation may be to ensure that the national best practice utilities are also benchmarked .

There are some additional problems with using an international data set for benchmarking. The main problem is that of comparability of data. One is forced to use the strategy of the least common denominator. A special issue is the correct handling of currency exchange rates. There are really only two practical alternatives; the average rates of exchange and the Purchasing Power Parity (PPP) as measured by OECD. The latter approach is chosen here. Relative differences in input prices like wage rates and rates of return on capital may also create problems as to distinguish between substitution effects and inefficiency.

[1] This essay was published in Resource and Energy Economics in 2003.

According to the findings in Jamasb and Pollitt (2001) international comparisons are often restricted to comparison of operating costs because of the heterogeneity of capital. As a precondition for international comparisons they focus on improving the quality of the data collection process, auditing, and standardization within and across countries. Our data have been collected specifically for this study by national regulators, and special attention has been paid to standardize the capital input as a replacement cost concept.

When doing international benchmarking for the same type of production activity in several countries, applying a common frontier technology seems to yield the most satisfactory environment for identifying multinational peers and assessing the extent of inefficiency. In our exercise for a sample of large electricity distribution utilities from Denmark, Finland Norway, Sweden and the Netherlands it is remarkable that peers come from all countries. The importance of exposing national units, and especially units that would have been peers within a national technology, to international benchmarking is clearly demonstrated. The multinational setting has called for the development of new indices to capture the cross-country pattern of the nationality of peers and the nationality of units in their referencing sets. Bilateral Malmquist productivity comparisons can be performed between units of particular interest in addition to country origin, e.g. sorting by size, or location of utility (urban - rural), etc. We have focused on a single unit against the (geometric) average performance of all units, as well as bilateral comparisons of (geometric) averages of each country. Our results point to Finland as the most productive country within the common technology. This result reflects the more even distribution of the Finnish units and the high share of units above the total sample mean of efficiency scores.

Essay II. Far out or alone in the crowd: classification of self evaluators in DEA

The DEA method classifies units as efficient or inefficient. The units found strongly efficient in DEA studies on efficiency can be divided into self-evaluators and active peers, depending on whether the peers are referencing any inefficient units or not. Self-evaluators was introduced by Charnes et al. (1985). The contribution of the paper starts with subdividing the self-evaluators into interior and exterior ones. The exterior self-evaluators are efficient “by default”; there is no firm evidence from observations for the classification. Self-evaluators may most naturally appear at the “edges” of the technology, but it is also possible that self-evaluators appear in the interior. It may be of importance to distinguish between the self-evaluators being exterior or interior. Finding the influence of some variables on the level of efficiency by running regressions of efficiency scores on a set of potential explanatory variables is an approach often followed in actual investigations. Using exterior self-evaluators with efficiency score of 1 in such a “two-stage” procedure may then distort the results, because to assign the value of 1 to these self-evaluators is arbitrary. Interior self-evaluators, on the other hand, may have peers that are fairly similar. They should then not be dropped when applying the two- stage approach.

A method for classifying self-evaluators based on the additive DEA model, either CRS or VRS, is developed. The exterior strongly efficient units are found by running the enveloping procedure “from below”, i.e. reversing the signs of the slack variables in the additive model, after removing all the inefficient units from the data set. Which units of the strongly efficient units from the additive model that turn out to be self-evaluators or active peers, will depend on the orientation of the efficiency analysis, i.e. whether input-or output orientation is adopted. The classification into exterior and interior peers is determined by the strongly efficient units turning out to be exterior ones running the “reversed” additive model.

The exterior self-evaluators units should be removed from the observations on efficiency scores when performing a two-stage analysis of explaining the distribution of the scores. The application to municipal nursing- and home care services of Norway shows significant effects of removing exterior self-evaluators from the data when doing a two-stage analysis. Thus the conclusions as to explanations of the efficiency score distribution will be qualified taking our new taxonomy into use.

Essay III. Climbing the efficiency stepladder: robustness of efficiency scores in DEA

The robustness of the efficiency scores in DEA has been addressed in a number of research papers. There are several potential problems that can disturb precise efficiency estimation, such as sampling error, specification error, and measurement error. It is almost exclusively the latter that is dealt with in this paper.

It has been proven analytically that the DEA efficiency estimators are asymptotically consistent given that a set of assumptions is satisfied. The most critical assumption might be that there are no measurement errors. The DEA method estimates the production possibility set by enveloping the data as close as possible, in the sense that the frontier consists of convex combinations of actual observations, given that the frontier estimate can never be “below” an observed value. If the assumption of no measurement error is broken we might observe input-output vectors that are outside the true production possibility set, and the DEA frontier estimate will be too optimistic. Calculating the efficiency of a correctly measured observation against this optimistic frontier will lead to efficiency scores that are biased downwards. In other words, even symmetric measurement errors can produce efficiency estimates that are too pessimistic. It is of crucial importance for the practical use of the efficiency scores that information about their sensitivity is available.

The reason why measuring sensitivity is a challenge is in a sense related to the difficulty with looking at n-dimensional space. In two dimensions, and possibly three, one can get an idea of the sensitivity of one observation efficiency score by visually inspecting a scatter diagram. But when the number of dimensions is higher than three, help is needed. The Efficiency Stepladder method introduced in this paper is an offer to empirically oriented DEA applications.

This paper is not about detecting outliers; it is about investigating the robustness of each DMUs efficiency score. The main inspiration is Timmer (1971), and the intention is to offer a crude and simple method that works relatively quickly and is available to practitioners as a freely downloadable software package.

In the following only DEA related approaches are considered. There are mainly two ways sensitivity to measurement error in DEA has been examined: (1) perturbations of the observations, often with strong focus on the underlying LP model, and (2) exclusion of one or more of the observations of the dataset.

The Efficiency Stepladder is based on the latter alternative. The main idea is to examine how the efficiency score of a given inefficient DMU develops as the most influential other DMU is removed in each of the iterative steps. The first step is to determine which of the peers whose removal is associated with the largest increase in the efficiency score. This peer is permanently removed, and the DEA model is recalculated giving a new efficiency score and a new set of peers. The removal continues in this fashion until the DMU in question is fully efficient. This series of iterative DMU exclusions provides an “efficiency curve” of the increasing efficiency values connected with each step.

There are few alternative approaches available that provide information about the sensitivity of efficiency scores. Related methods in the literature are Peeling (Barr et al., 1994), Efficiency Order (Sinuany-Stern et al., 1994) and Efficiency Depth (Cherchye et al., 2000). Peeling consists of removing all the frontier units in each step. There are also similarities between the Efficiency stepladder and the Efficiency order/Efficiency Depth methods. The main difference is that the Efficiency stepladder approach is concerned with the stepwise increase in the efficiency scores after each iterative peer removal, while the Efficiency Order/Efficiency Depth methods are more concerned with the number of observation removals that is required for the DMU in question to reach full efficiency.

The empirical application is mainly used as an illustration on how the Efficiency stepladder method works on real world data. The application is used to show what kind of analysis can be performed using this method. To carry out a full scale empirical analysis is an extensive undertaking, and is outside the scope of this paper.

Ideally sensitivity analysis, detection of potential outliers, and estimation of sampling bias should be carried out simultaneously. It is easier to detect outliers if we have some information about the sampling bias, and it is easier to estimate sampling bias if we have first identified the outliers. There have been developments made on all these areas in the last few years, but at the time of writing no single method offers a solution to all the mentioned challenges.

The Efficiency stepladder method is simple and crude, but it can still be useful for applied DEA investigations. It should be thought of as one way safe: An Efficiency stepladder that is very steep is a clear indication that the DEA estimated efficiency is strongly dependent on the correctness of a low number of other observations. A slow increase on the other should not be interpreted as a strong indication that the efficiency is at least this low. The reason is that the method is only one-step-optimal. In addition to measuring the sensitivity of the e-scores for efficient and inefficient units, it might be used in combination with bootstrapping to identify possible outliers. The necessary software for carrying out the Efficiency stepladder calculations will be made available from the author's website.

The purpose of the ESL method is to examine the sensitivity of the efficiency scores for measurement errors. Bootstrapping on the other hand is in the DEA context (primarily) used to measure sensitivity to sampling errors. We would expect that a DMU with a large ESL(1) value would also have a large standard error of the bias corrected efficiency score. The reason is that we expect the part of the (input, output) space where the DMU is located to be sparsely populated.

Tentative runs have shown statistically significant and positive correlation between the ESL(1) values and the standard errors of the bootstrapped bias corrected efficiency scores. Furthermore, there is strong empirical association between the ESL(1) values for the fully efficient DMUs (=superefficiency) and the sampling bias estimated using bootstrapping. This is a promising topic for further research.

Essay IV. Efficiency of Norwegian construction firms

Low productivity growth of the construction industry in the nineties (based on national accounting figures) is causing substantial concern in Norway. To identify the underlying causes investigations at the micro level are needed. However, efficiency studies at the micro level of the of the construction industry are very rare.

The objective of this study is to analyze productive efficiency in the Norwegian construction industry. A piecewise linear frontier is used, and technical efficiency measures (Farrell, 1957) are calculated on cross section data following a DEA (data envelopment analysis) approach (Charnes et al., 1978).

The DEA efficiency scores are bias corrected by bootstrapping (Simar and Wilson, 1998, 2000), and a bootstrapped scale specification test is performed (Simar and Wilson, 2002). A new contribution is to use weights based on the standard errors from the bootstrapped bias correction in the two stage model when searching for explanations for the efficiency scores.

One reason for the small number of efficiency analyses of the construction industry may be the problem to “identify” the activities in terms of technology, inputs and outputs in this industry. It is well known that there are large organizational and technological differences between building firms. Even when the products are seemingly similar there are large differences in the way projects are carried out. For instance some building projects use a large share of prefabricated elements, while other projects produce almost everything on the building site. This often happens even when the resulting construction is seemingly similar. It is interesting to note that projects with such large differences in the technological approach can exist at the same time. Moreover, the composition of output varies a lot between different construction companies so the definition of the output vector may also be a problem. Thus to capture such industry characteristics, a multiple input multiple output approach is required.

Large differences in the efficiency and productivity scores were discovered. One important lesson that can be learned from this application is the danger of taking the efficiency scores from uncorrected DEA calculations at face value. If one decided to learn from a few DMUs based on their uncorrected efficiency scores, one might get into trouble. It is not unreasonable to think that similar things have happened in the last few years as DEA has been embraced by a very large number of practitioners (researchers and consultants).

It would be interesting if the large number of empirical DEA papers were recalculated using the bootstrap methodology. Anecdotal observations indicate that very few practitioners use bootstrapping. The reason for this might be that bootstrapping is not yet available in the standard DEA software packages.

Based on a scale specification test, a variable returns to scale specification was selected. A scale chart indicated that firms with total production values lower than 100 mill. NOK might be operating at a suboptimal scale level.

The differences in the efficiency scores may be explained by environmental and managerial variables. Such variables have been tried in a two stage approach. A new contribution is the demonstration of how one can use the standard errors from the bias correction in stage one to improve the power of the regression model in stage two.

Five possible explanations were examined for empirical relevance, and four of them were found to be statistically significant in a multivariate weighted regression setting. More detailed data would be necessary before strong conclusions can be made, but there are indications that the most efficient building firms are characterized by high average wages, low numbers of apprentices, diversified product mixes and high numbers of hours worked per employee.

References

Barr, R.S., M.L. Durchholz and Seiford, L., 1994, Peeling the DEA Onion. Layering and Rank-Ordering DMUs Using Tiered DEA, Southern Methodist University technical report, Dallas, Texas.

Caves, D.W., L.R. Christensen and E. Diewert, 1982a, The economic theory of index numbers and the measurement of input, output, and productivity, *Econometrica* 50, 1393-1414.

Charnes, A., Cooper, W.W. and Rhodes, E., 1978, Measuring the efficiency of decision making units, *European Journal of Operations Research* 2, 429-444.

Charnes, A., Cooper, W.W., Lewin, A.Y. , Morey, R.C., and Rousseau, J.J.., 1985. Sensitivity and Stability Analysis in DEA. *Annals of Operations Research* 2 139-150.

Cherchye, L. Kuosmanen, T. and Post, G.T., 2000, New Tools for Dealing with Errors-In-Variables in DEA, Katholike Universiteit Leuven, Center for Economic Studies, Discussion Paper Series DPS 00.06.

Farrell, M.J.,1957, The measurement of productive efficiency, *J.R. Statis. Soc*. Series A 120, 253-281.

Førsund, F. R. and S. A. C. Kittelsen, 1998, Productivity development of Norwegian electricity distribution utilities, *Resource and Energy Economics* 20(3), 207-224.

Jamasb, T. and M. Pollitt, 2001, Benchmarking and regulation: international electricity experience, *Utilities Policy* 9(3), 107-130.

Sinuany-Stern, Z., A. Mehrez and A. Barboy, 1994, Academic Departments Efficiency via DEA, *Computers Ops. Res*., vol. 21, No. 5, pp. 543-556.

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., 2000, A general methodology for bootstrapping in nonparametric frontier models, *Journal of Applied Statistics* 27, 779--802.

Simar, L. and Wilson, P., 2002, Nonparametric Tests of Returns to Scale, *European Journal of Operational Research*, 139, 115-132

Timmer, C.P., 1971, Using a Probibalistic Frontier Production Function to Measure Technical Efficiency, *Journal of Political Economy*, Vol. 79, No. 4 (Jul. – Aug. 1971), 776-794.

INTERNATIONAL BENCHMARKING OF ELECTRICITY DISTRIBUTION UTILITIES[*]

by

Dag Fjeld Edvardsen

The Norwegian Building Research Institute *Forskningsvn. 3 b*
P.O. Box 123, Blindern, 0314 Oslo, Norway

and

Finn R. Førsund[±]

Department of Economics, University of Oslo, and the Frisch Centre
P.O. Box 1095, Blindern, 0317 Oslo, Norway

Abstract: Benchmarking by means of applying the DEA model is appearing as an interesting alternative for regulators under the new regimes for electricity distributors. A sample of large electricity distribution utilities from Denmark, Finland, Norway, Sweden and the Netherlands for the year 1997 is studied by assuming a common production frontier for all countries. The peers supporting the benchmark frontier are from all countries. New indices describing cross-country connections at the level of individual peers and their inefficient units as well as between countries are developed, and novel applications of Malmquist productivity indices comparing units from different countries are performed.

Key words: Electricity distribution utility, benchmarking, efficiency, DEA, Malmquist productivity index

JEL classification: C43, C61, D24, L94.

[*] The study is done within the research project "Efficiency in Nordic Electricity Distribution" at the Frisch Centre, financed by the Nordic Economic Research Council. Finn R. Førsund was visiting fellow at ICER during the fall 2001 and spring 2002 when completing the paper. We are indebted to a group of Danish, Dutch, Finnish, Norwegian and Swedish electricity regulators for cooperation and comments on earlier drafts at project meetings in Denmark, Norway, Finland and the Netherlands. We will especially thank Susanne Hansen, Kari Lavaste and Victoria Shestalova for written comments. We are indebted to Sverre A. C. Kittelsen for valuable comments on the last draft, and a referee for stimulating further improvements.

The electricity regulators, headed by Arne Martin Torgersen and Eva Nœss Karlsen from NVE, have done extensive work on data collection. However, notice that the responsibility for the final model choice and focus of the study rests with the authors. Furthermore, the analysis is only addressing technical efficiency measurement, and in particular not cost efficiency. The study is not intended for regulatory purposes.

[±] Corresponding author. Tel.:+47-2285-5132; fax: +47-2285-5035
Email address: f.r.forsund@econ.uio.no (F.R. Førsund).

1. Introduction

Improvement of efficiency in electricity distribution utilities has come on the agenda, as an increasing number of countries moved towards deregulation of the sector in the last decade. A key element in assessing potentials for efficiency improvement is to establish *benchmarks* for efficient operation. A standard definition of benchmarking is a comparison of some measure of actual performance against a reference performance. One way of obtaining a comprehensive benchmarking as opposed to partial key ratios is to establish a frontier production function for utilities, and then calculate efficiency scores relative to the frontier.

In this study a piecewise linear frontier is used, and technical efficiency measures (Farrell, 1957) and Malmquist productivity measures (Caves et al., 1982a) are calculated by employing the DEA model (Charnes et al., 1978). The DEA model has been used in several studies of the utilities sector recently (see a review in Jamasb and Pollitt, 2001). A special feature of the present cross section study is that the data (for 1997) is based on a sample of utilities from five countries: Denmark, Finland, The Netherlands, Norway and Sweden. Most of the efficiency studies of utilities have been focusing on utilities within a single country (Førsund and Kittelsen, 1998), but a few studies have also compared utilities from different countries (Jamasb and Pollitt, 2001). In some cases an international basis for benchmarking is a necessity due to the limited number of similar firms within a country. When the number of units is not the key motivation for an international sample for benchmarking, the motivation may be to ensure that the national best practice utilities are also benchmarked[1].

There are some additional problems with using an international data set for benchmarking. The main problem is that of comparability of data. One is forced to use the strategy of the least common denominator. A special issue is the correct handling of currency exchange rates. There are really only two practical alternatives; the average rates of exchange and the Purchasing Power Parity (PPP) as measured by OECD. The latter approach is chosen here. Relative differences in input prices like wage rates and rates of return on capital may also create problems as to distinguish between substitution effects and inefficiency.

[1] An alternative is to use hypothetical units based on engineering information, as mentioned already in Farrell (1957). In Chile and Spain hypothetical model best practice units are used for benchmarking (Jamasb and Pollitt, 2001).

According to the findings in Jamasb and Pollitt (2001) international comparisons are often restricted to comparison of operating costs because of the heterogeneity of capital. As a precondition for international comparisons they focus on improving the quality of the data collection process, auditing, and standardization within and across countries. Our data have been collected specifically for this study by national regulators, and special attention has been paid to standardize the capital input as a replacement cost concept.

Regarding the extent of international studies Jamasb and Pollitt (2001) found that 10 of the countries covered in the survey (OECD- and some non-OECD countries) have used some form of benchmarking, and about half of these use the frontier-oriented methods: DEA, Corrected Least Squares (COLS) and the Stochastic Frontier Approach (SFA). They predict that benchmarking is likely to become more common as more countries implement power sector reforms. (For an opposing view, see Shuttleworth, 1999.)

The rest of the paper is organized in the following way: In Section 2 the DEA model is introduced and new indices are developed to capture the cross-country pattern of the nationality of peers and the nationality of units in their sets of associated inefficient units. Malmquist productivity approaches are developed for cross section international comparisons. In Section 3 the theory of distribution of electricity as production is briefly reviewed with regards to the choice of variable specification. Structural differences between the countries revealed by the data are illustrated. The results on efficiency distributions and inter-country productivity differences using Malmquist indices are presented in Section 4. Conclusions and further research options are offered in Section 5.

2. The methodological approach

2.1. The DEA model

As a basis for benchmarking we will employ a piecewise linear frontier production function exhibiting the transformations between outputs, y_m (m = 1,..,M) and the substitutions between inputs, x_s (s = 1,..,S). We will assume constant returns to scale (CRS). The frontier is enveloping the data as tightly as possible, and observed utilities, termed best practice, will form the benchmarking technology. The Farrell technical efficiency measures are calculated simultaneously with determining the nature of the envelopment, subject to basic properties of the general transformation of inputs into outputs (Färe and Primont, 1995). The efficiency scores for the input oriented DEA model, E_i for utility no i ($i \in N$ = set of units) are found by solving the following linear program:

$$\begin{aligned}
&E_i = Min \quad \theta_i \\
&s.t. \\
&\sum_{j \in N} \lambda_{ij} y_{mj} - y_{mi} \geq 0 \quad , \quad m = 1,..,M \\
&\theta_i x_{si} - \sum_{j \in N} \lambda_{ij} x_{sj} \geq 0 \quad , \quad s = 1,..,S \\
&\lambda_{ij} \geq 0 \quad , \quad j \in N
\end{aligned} \qquad (1)$$

The point $(\sum_{j \in N} \lambda_{ij} x_{1j},..,\sum_{j \in N} \lambda_{ij} x_{Sj},\sum_{j \in N} \lambda_{ij} y_{1j},..,\sum_{j \in N} \lambda_{ij} y_{Mj})$ is on the frontier and is termed the *reference point*. In the CRS case the input- and output oriented scores are identical. However, we may need to keep non-discretionary variables fixed when calculating the efficiency scores. Then, in the case of an output fixed, the input-oriented model (1) and the scores remain the same. But if one of the inputs is fixed the efficiency correction of that input constraint in (1) is dropped and the numerical results for efficiency scores may be different.[2]

2.2. The Peers

The efficient units identified by solving the problem (1) are defined as *peers* if the efficiency score is 1 and all the output- and input constraints in (1) are binding. Each inefficient unit will be related to one or more benchmark or peer units. Let *P* be the set of peers and *I* the set of inefficient units, $P \cup I = N$. A *Reference set* or *Peer group set* for an inefficient unit, *i*, (Cooper, Seiford and Tone, 2000), is defined as:

$$P_i = \left\{ p \in P : \lambda_{ip} > 0 \right\} \quad , \quad i \in I \qquad (2)$$

Each inefficient unit, *i*, has a positive weight, λ_{ip}, associated with each of its peers, *p*, from the solution of the DEA model (1). The weights, λ_{ip}, are zero for inefficient units not having unit *p* as a peer. Since all peers have the efficiency score of one there is a need to discriminate between peers as to importance as role models. Measures used in the literature are a pure

[2] Correspondingly, an output-oriented model will be different if one of the outputs is fixed (but not if one of the inputs is fixed), since the constraint involving this variable will be reformulated to hold without the efficiency correction of the output variable for the unit being investigated.

count measure based on the number of peer group sets (2) that a peer is a member of, calculating a *Super-Efficiency* measure for a peer against a frontier recalculated without this peer in the data set supporting the frontier (Andersen and Petersen, 1993), and a *Peer index* (Torgersen et al., 1996) showing the importance of a peer as a role model based on the share of the input savings of the inefficient units referenced by a peer, weighted by the weights λ_{ip} found by solving (1).

2.3. Cross group influence of peers

For our situation with units from different countries we are more interested in developing measures that show the interconnections between peers and inefficient units from different countries. We will need to consider a peer and the set of inefficient units that are referenced by the peer. We will term this apparently new set in the literature, I_p, the *Referencing set* for a peer, p:

$$I_p = \left\{ i \in I : \lambda_{ip} > 0 \right\} \ , \quad p \in P \tag{3}$$

One approach is to focus on the country distribution of the inefficient units in a peer's referencing set. Units must now be identified by country. Let L be the set of countries and I^q the set of inefficient units of country q ($\underset{q \in L}{\cup} I^q = I$). Partitioning the Referencing set (3) by grouping the inefficient units according to country yields:

$$I_p^q = \left\{ \ i \in I^q : \lambda_{ip} > 0 \right\} \quad , p \in P, q \in L \ , \ \underset{q \in L}{\cup} I_p^q = I_p \tag{4}$$

Let the number of units in the Referencing set (3) be $\#I_p$, the number of units in the set (4) be $\#I_p^q$ and the set of peers from country q be P^q ($\underset{q \in L}{\cup} P^q = P$). The *Degree of peer localness* index, DL_p^q, for peer p in country q, is then defined as:

$$DL_p^q = \frac{\#I_p^q}{\#I_p} \quad , \quad p \in P^q, \quad q \in L \tag{5}$$

The index varies between zero and one. Zero means that the peer is “extreme- international”, only referencing inefficient units from other countries, and one means that the peer is “extreme- national”, only referencing inefficient units from own country.

In Schaffnit et al. (1997) a count measure was developed describing the number of inefficient units belonging to a group referenced by peers from another group, relative to the total number of units of the first group (may be the number of *inefficient* units would be more appropriate). In order to obtain more detailed information we will instead develop an index for *Cross-country peer importance* by using characteristics of the inefficient units analogous to the Peer index mentioned above. In the case of input orientation[3] the index, ρ^s_{qr}, can be established by weighing the saving potential of an input, *s*, for the inefficient units from a country, *q* ($= x_{ks}(1-E_k)$ $, k \in I^q$), with the relevant λ_{kp}- weights associated with peers from another country, *r* ($p \in P^r$), being in the peer group set of the inefficient units from country *q*, and then comparing with the total saving potential of all inefficient units in country *q*:

$$\rho^s_{qr} = \frac{\sum_{p \in P^r} \sum_{k \in I^q} (\lambda_{kp} / \sum_{p' \in P} \lambda_{kp'}) x_{ks}(1-E_k)}{\sum_{k \in I^q} x_{ks}(1-E_k)} \quad , \quad s = 1,..,S, \quad q,r \in L \tag{6}$$

The weights in the numerator are normalized with the sum of weights for all peers for the inefficient unit *k* from country *q*. In the variable returns to scale case this sum is restricted to be one, but not in the CRS case we are working with. This index will be input (output) variable specific, as is the case for the Peer index. The maximal value of the index is 1. This will be the case if peers belonging to country *r* reference all the inefficient units of country *q*, and that they are not referenced by peers from any other country. The minimal index value of zero is obtained if peers from country *r* do not reference any inefficient unit from country *q*.

2.4. The Malmquist productivity index

The Malmquist productivity index, introduced in Caves et al. (1982a), is a binary comparison of the productivity of two entities, usually the same unit at different points in time, but we may also compare different units at the same point in time. Let the set of units in country *q* be N^q, etc. ($\underset{q \in L}{\cup} N^q = N$). The output- and input vectors of a unit, *j*, are written $y_j = (y_{j1},..,y_{jM}), x_j = (x_{j1},..,x_{jS}), \ j \in N$. The Malmquist productivity index, $M^q_{k,l}$, for the two units *k* and *l* from country *q* and *r* respectively, is:

$$M^q_{k,l}(y_k, x_k, y_l, x_l) = \frac{E^q_l(y_l, x_l)}{E^q_k(y_k, x_k)} \quad , \quad k \in N^q, \quad l \in N^r, \quad q \in L \tag{7}$$

The Malmquist index is the ratio of the Farrell technical efficiency measures for the two units, as calculated by solving the program (1). The superscript on the indexes shows the

[3] An output-oriented Cross-country peer index can be formulated analogously following the definition of the Peer index in Torgersen et al. (1996) for output orientation.

reference technology base (relevant for one of the units being compared, i.e. *q* means that the efficiency measures are calculated with respect to the frontier for country *q*). We follow the convention of having the unit indicated first in the subscript of the Malmquist index on the lhs of (7) in the denominator and the second in the numerator, thus unit *l* is more productive than unit *k* if $M_{k,l}^{q} > 1$, and vice versa. If it is appropriate to operate with different reference technologies for countries, following Färe et al. (1994) the Malmquist index can be decomposed multiplicatively into a term reflecting each unit catching up with its reference technology, and a term reflecting the distance between the two reference technologies.[4]

Since we are dealing with countries it may also be of interest to compare productivity levels between countries. The crucial point concerning how to construct indices for comparisons is the assumption about production technologies. There are two basic alternatives:

i) A common frontier technology may be assumed, allowing utilities from different countries to support the DEA envelope.

ii) The technologies are national, i.e. only own country units may be best practice ones.

2.5. Common inter- country technology

As pointed out in Caves et al. (1982b) it is an advantage to use a circular index when comparing productivities of two countries (units). Berg et al. (1992), (1993), and Førsund (1993) demonstrate that the Malmquist index (7) is not circular (see also the general discussion in Førsund, 2002). In the case of the same frontier technology being valid for all countries, corresponding to assumption i) above, the index is then circular. The calculation of the Malmquist productivity index is greatly simplified, since the benchmark technology will be common for all productivity calculations. The notation of the expressions below is simplified by removing the technology index.

A useful characterization of the productivity of a unit, *k*, may be obtained by comparing the efficiency score for this unit with the geometric mean of all the other scores, following up Caves et al. (1982b), (p. 81, Eq. (34)), where the productivity of one unit was measured against the geometric mean of the productivities of all units:

[4] An application of such decomposition in a study of Norwegian electricity distributors is found in Førsund and Kittelsen (1998).

$$\bar{M}_k = \frac{E_k(y_k, x_k)}{\left[\Pi_{l\in N} E_l(y_l, x_l)\right]^{1/\#N}} \quad , \quad k \in N \tag{8}$$

where *#N* is the total number of all utilities. This geometric mean-based Malmquist index is a function of all observations. To focus on bilateral productivity comparisons between countries, one way of formulating this is to compare the geometric means of efficiencies over units for each country, *q* and *r*, symbolized by the sub-index *g(r,q)*:

$$\bar{M}_{g(r,q)} = \frac{\left[\prod_{k\in N^q} E_k(y_k, x_k)\right]^{1/\#N^q}}{\left[\prod_{l\in N^r} E_l(y_l, x_l)\right]^{1/\#N^r}} \;, \; q, r \in L \tag{9}$$

where #N^q and #N^r are the total number of utilities within country *q* and *r* respectively. This geometric mean-based Malmquist index is a function of all the observations in countries *r* and *q*. The index may be termed the *bilateral country productivity index*, and is circular, in the sense that the index is invariant with respect to which third country efficiency score average we may wish to compare with countries *q* and *r*.

If we want to express how, on the average, the units within a country, *q*, are doing compared with the average over all units, the country *r* specific index in the denominator of (9) can be substituted with the geometric average of the efficiency scores of all the utilities, i.e. the denominator in (8). The geometric mean of efficiencies for units within a country, symbolized by the sub index *g(q)*, is compared with the geometric mean over all units:

$$\bar{M}_{g(q)} = \frac{\left[\prod_{k\in N^q} E_k(y_k, x_k)\right]^{1/\#N^q}}{\left[\prod_{l\in N} E_l(y_l, x_l)\right]^{1/\#N}} \quad , \quad q \in L \tag{10}$$

3. **Model specification and data**

3.1. Distribution as production

In the review of transmission and distribution efficiency studies Jamasb and Pollitt (2001) point to the variety of variables that have been used as an indication that there is no firm consensus on how the basic functions of electric utilities are to be modeled as production activities. However, they mention that this may, to some extent, be explained by the lack of data.

Modeling the production activity of transportation of electricity has old traditions within engineering economics (see e.g. Førsund (1999) for a review). On a general abstract level the outputs of distribution utilities are the energy delivered through a network of lines and transformers to the consumption nodes of the network and losses in lines and transformers. The inputs are the energy received by the utility, real capital in the form of lines and transformers, and labor and materials used for general distribution activities. Due to the high number of customers for a standard utility it is impossible to implement the conceptualization of a multi-output production function to the full extent. The usual approximation is to operate with total energy delivered and number of customers separately as outputs (Salvanes and Tjøtta, 1994). The latter variable is also often used in engineering studies as the key dimensioning output variable, and taken as the absolute size of a utility (Weiss, 1975). In engineering studies the load density may be a characterization of capital. Load density is the product of customer density and coincident peak load per customer (kWh per square mile). The maximum peak load may also describe capital as a quality attribute, or be used as an output attribute characterizing energy delivered.

In the short run the utilities take the existing lines, transformer capacity and the geographical distribution and number of customers as given. But, as pointed out in Neuberg (1977), this is not the same as saying that these variables must be regarded as constants in our analysis. Past decisions reflected in configurations of lines and transformers may give rise to current differences in efficiency. These variables that are exogenous for the firm, may be seen as endogenous from the point of view of society. Even distribution jurisdictions can be rearranged, making the number of customers endogenous.

The role of lines varies. It can be regarded as a capital input, but it is also used as a proxy for the geographical extent of the service area. For fixed geographical distribution of customers the miles of distribution line would be approximately set (but note the possibilities of inefficient configurations), thus line length may serve as a proxy for service area. The service area can be measured in different ways. The problem is to find a measure the utility cannot influence (see Kittelsen (1993) and Langset and Kittelsen, 1997). Due to probability of wireoutage and cost of servicing the extent of customer area will influence distribution costs. Non-traditional variables such as size of service area may also be used to specify differences in the production system or technology from utility to utility.

According to the extensive review in Jamasb and Pollitt (2001) the most frequently used inputs are operating costs, number of employees, transformer capacity, and network length. The most widely used outputs are units of energy delivered, number of customers, and size of service area.

3.2. Choice of model specification

Concerning our choice of input variables it has not been possible to use a volume measure of labor due to the lack of this information for one country (Denmark). Instead a cost measure has been adopted. Labor cost, other operating costs and maintenance have been aggregated to total operating and maintenance costs (TOM). We then face the problem mentioned in the introduction about national differences in wages for labor. It has been chosen to measure TOM in Swedish (SEK) prices.

A measure for real capital volume has been established for 1997 by the involved regulators by first creating for the sample utilities a physical inventory of existing real capital in the form of length of types of lines (air, under ground and sea) distributed in three classes according to voltage, categories of transformers according to type (distribution, main) and capacity in kV, transformer kiosks for distribution, and transformer stations for main transformers. The number of capital items for each country has been in the range of 60 to 100. As a measure of real capital the *replacement value* (RV) is the theoretically correct measure (Johansen and Sørsveen, 1967). To obtain such a measure, aggregation over the categories has been necessary due to the large number of items. The same weights should be used, i.e. using national prices will not yield a correct picture if prices differ. It has been chosen to use Norwegian prices for all countries. A more preferred set of weights may be average prices for all countries, but it has not been feasible to establish such a database for this study. Although lines and transformers have been used separately as inputs in the literature (see e.g. Hjalmarsson and Veiderpass (1992a), (1992b) and Jamasb and Pollitt, 2001), the groups have been aggregated into a single aggregated capital volume measure in this study, partly due to different classification systems used by the countries.

We will simplify on the energy input side and only use the loss in MWh in the system as a proxy for input. This variable will also capture a quality component of the distribution system. A problem is that data on losses may be measured with less precision due to measuring periods not coinciding with the calendar year. For some countries an average loss for the last three years is used, while loss for the last year or its estimate is used for other countries.

On the output side energy delivered and the number of customers are used as outputs. The countries have information on low and high voltage, but since the classification of high and low voltage differs, we had to use the aggregate figures. Some measure of geographical configuration of the distribution networks should also be included for a relevant analysis of efficiency. In this study the total length of distribution lines is the only available measure for service area. In addition to service area the density of customers of a distributor is usually considered to influence the efficiency. But when using absolute number of customers and energy delivered as separate outputs there is no room for an additional density variable of the type energy per customer. By nature of the radial efficiency measure, the reference point on the frontier has the same energy-per-customer density as the observation in question. The countries involved have very different population densities. But it is not so obvious how this will influence efficiency in distribution. A rural distributor in Norway may serve a community located along a valley bottom with people living fairly close to each other, while the geographical area of the municipality may include vast uninhabited area of mountains and forests above the valley floor. A densely populated area in the Netherlands may not necessarily save on lines per unit of area if low-rise housing dominates.

3.3. The data structure

An overview of key characteristics of the data is presented in Table 1. The difference in size between utilities is large, as revealed by the last two columns. A summary of the structure of the data of the individual countries is shown in the radar diagram in Figure 1, where country averages relative to the total sample averages (the 100% contour) are portrayed. By using the contour curves for percentages, relative comparisons can also be done between countries. The domination in size of the Netherlands is obvious in all dimensions except for energy delivered. The Netherlands is especially large in number of customers, but also in replacement value. It is relatively smaller in length of lines. Norway is

Table 1 Summary statistics. Cross-section 1997. Number of units 122

	Average	Median	Standard Deviation	Minimum	Maximum
TOM(kSEK)	152388	97026	182923	11274	981538
LossMWh	91449	52318	104777	7020	615281
RV (kSEK)	2826609	1907286	3288382	211789	22035846
NumCust	109260	55980	163422	20035	1052096
TotLines	7640	4948	8824	450	54166
MWhDelivered	2110064	1003472	2815025	166015	178054730

largest with respect to energy delivered and also correspondingly large in energy loss, although with a smaller value than the Netherlands. Sweden stands out with relatively high

operating and maintenance costs (TOM), while Finland stands out with a high number for length of lines. Denmark has the smallest number for length of lines and energy loss, and has a relatively high number of customers. The combinations of number of customers and length of line show the highest customer density in the Netherlands and then Denmark second, and the lowest density in Finland.

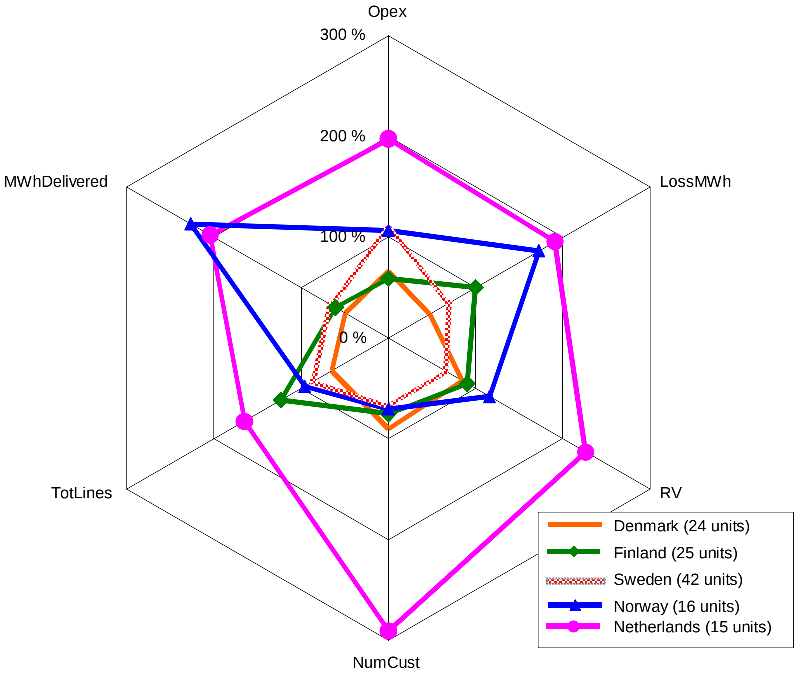

Figure 1. The average structure of the countries

4. The results

4.1. Efficiency scores

The distribution of efficiency scores[5] for model (1) is shown in Figure 2. The units for each country are grouped together and sorted according to ascending values of the efficiency score. Each bar represents a unit, an electricity distribution utility company. The size of each unit, measured as total operating and maintenance costs (TOM) (including labor costs), is proportional to the width of each bar.[6] The efficiency score is measured on the vertical axis and the TOM values measured in SEK (in 1000) are accumulated on the horizontal axis. As a

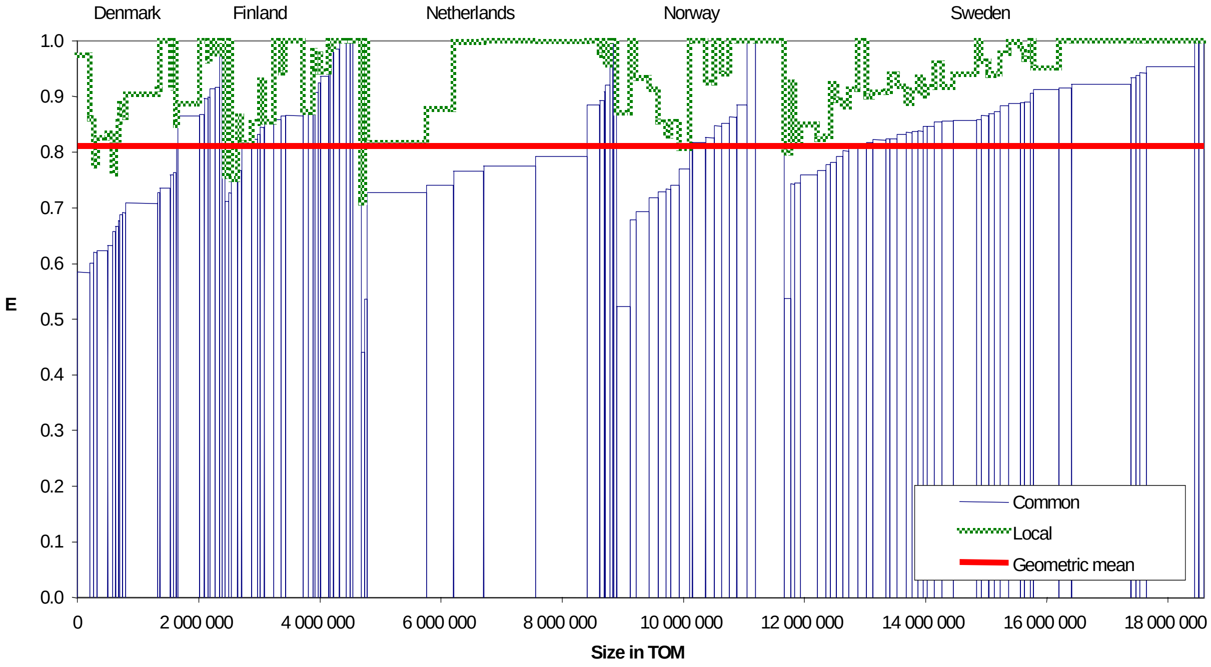

Figure 2. Country distribution of efficiency scores

[5] The efficiency score values are given in Edvardsen and Førsund (2002). One Dutch unit has been removed from the original data set after performing a sensitivity test and considering the atypical structure of the unit. Notice that service area may be regarded as a fixed non-discretionary variable without any consequence for the values of the efficiency scores since input-orientation is adopted, cf. the discussion below the model (1).

[6] The regulators chose this input-based size measure as being most relevant to them. Other candidates for size variables are mentioned in Section 3. It does not matter much, which one is chosen for the purpose of getting information about the location of units according to size.

general characterization the units are distributed in the interval from 0.44 to 1, and the share of TOM of fully efficient units is rather small, representing about 5 percent of accumulated TOM.

When looking at the country distributions it is remarkable that all countries have fully efficient units. This supports the use of a common technology, in the sense that no country is completely dominated by another, and all countries contribute to spanning the frontier. There are two aspects that the figure sheds light on: the size of the efficient units – measured by the input total operating costs – and how the efficient units stand out in the country specific distributions. For the three countries Denmark, the Netherlands and Sweden, the efficient units are quite small compared to average size within each country. This is especially striking for the Netherlands with the most pronounced dichotomy in size with one group of large units and the other with considerably smaller ones. The units within the group of large units have about equal efficiency levels, while the group with small units has units both at the least efficient part and the most efficient part of the distribution. The least efficient units have only half the value of the efficiency score than the average. For Finland and Norway the efficient units are closer to the medium size (disregarding the large Norwegian self-evaluator). The Swedish distribution is characterized by an even distribution of efficiency scores with large units being at the upper end of the inefficiency distribution, and medium- and small sized units being evenly located over the entire distribution.

The inefficient units with the highest efficiency scores have values quite a bit lower than 1 for Denmark, the Netherlands and Norway, while the values are much closer to the fully efficient ones in Finland and Sweden. The Norwegian distribution has no marked size pattern, but has a much more narrow range of the efficiency scores for the inefficient units than for Sweden. The range of the distribution for Finland is the narrowest without one or two extremely inefficient units like the case for the Netherlands, Norway and Sweden. Both for Finland and Denmark the largest units are located centrally in the distributions.

A rough measure of the total potential improvement for each country may be read off graphically in Figure 2 by the area between the value 1 for the efficiency score and the top of the bars representing the individual units for each country. The total savings potential for operating and maintenance costs is about 20 per cent (the potential for the other two inputs cannot be seen so accurately since TOM is used as the size variable). Finland has the smallest potential while Sweden and the Netherlands have the highest. As a summary expression for the different shapes of the efficiency distributions, different number of units and absolute size between units and location of size classes within country distributions, the country share of the savings potential $(=\sum_{i\in I^q} x_{is}(1-E_i)/\sum_{j\in I} x_{js}(1-E_j), q\in L)$ for each of the three inputs are set out in Table 2, using the radial projections.[7] Due to the large, inefficient Dutch units that we see in Figure 2, the Netherlands has a higher savings potential than the other countries, especially for replacement value of capital. Sweden has a high potential for total operating- and maintenance costs savings, and Norway for savings in energy loss. Denmark comes second to the Netherlands in saving potential for replacement value of capital, and has the smallest share for energy loss, roughly on the same level as Finland. Finland has significantly lower savings potential for total operating- and maintenance costs and replacement value of capital than the other countries.

In order to assess the efficiency of countries measuring an individual unit against the total (geometric) mean was introduced in Equation (8). The line of this geometric mean is inserted in Figure 2 ($\bar{E}$ = 0.82). The figure gives a visual impression of such comparisons. As overall characterizations we may note that the median efficiency score of Denmark and Norway is below the total mean, while the median value of Finland, the Netherlands and Sweden are higher. The Netherlands is a special case since all the large units are less productive than the sample (geometric) average.

4.2. Structural features of best- and worst practice units

From the efficiency distribution shown in Figure 2 we identify the 12 active peers (excluding the self-evaluator) and the 12 worst practice units and calculate the average input- and output values. Since we have 121 units this number represents the upper and lower deciles of the

Table 2. Country distribution of savings potential shares

	TOM	**LossMWh**	**RV**
Denmark	0.19	0.14	0.22
Finland	0.08	0.14	0.10
Netherlands	0.29	0.28	0.33
Norway	0.16	0.25	0.18
Sweden	0.28	0.19	0.17

[7] If the reference points on the frontier had been used some differences in shares may occur if slacks on the input constraints in (1) are present and unevenly distributed on countries.

distribution. The comparison is shown in Figure 3. It is the relative position in the radar diagram that reveals the structure. Both the best practice units (BP) and the worst practice units (WP) are smaller than the sample average (the 100% contour), except the input RV for the WP units. The BP units have, on average, higher values for all outputs than the WP units, and relatively lower number of customers compared with the WP units. Concerning inputs, the WP units have a significant over-use of capital (measured by the replacement value) leading to a much higher use of this input than for BP units, and also higher for total operating and maintenance costs (TOM), while energy loss is actually a little lower than for BP units.

4.3. The degree of localness of peers

We have already seen in Figure 2 that peers are found in all countries. An interesting question is the nature of the peers: are they multinational or pure national peers? If all the peers turn out to be national, the common technology is partitioned into country parts, and there is no foundation for international benchmarking. We will use the *Degree of peer localness index* (5) to describe the connection between a peer and its associated inefficient units, and the scope for international benchmarking. The information is found by partitioning the Referencing sets (3) on countries according to (4). This is done in the columns of Table 3 for

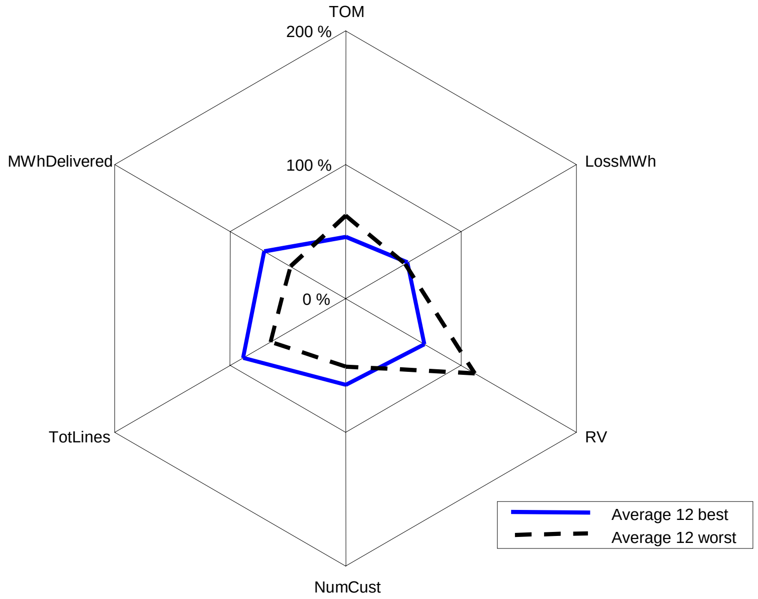

Figure 3. Structural comparison of best- and worst practice units]

each peer, entered according to nationality. An inefficient unit may appear in one or more of the peer columns.[8] All the active peers are referencing one or more inefficient units from their own country. We use as a criterion for a national peer that 50 percent or more of the inefficient units in its Referencing set are from its own country. The Degree of localness index in the last row of Table 3 shows that three peers are national. Both the two Swedish units (5022 and 5047), and one Danish unit (1023) have national roles as peers. The two Swedish peers have the highest Degree of peer localness index values of all peers, 1.00 and 0.73. A Finnish unit (2026) is close to being national, with an index value of 0.48. There are five truly multinational peers in the group of 13 efficient units, in the sense that they are referencing inefficient units from all five countries. Three of these stand out as referencing a considerably higher number of inefficient units, as seen from the second row from below. This is the pure count measure of peer importance. Only one peer (Swedish unit 5047) is truly national in the sense that it is only referencing inefficient units from its own country. Based on the pattern of country origin of peers and referenced units, Sweden has the most national peers with only one of two peers referencing a few inefficient units from Norway, Finland and Denmark. Denmark and Sweden seem to be furthest apart with reference to the common technology frontier, since two of Denmark's peers have only a single Swedish inefficient unit in their Referencing its sets, and only one Danish inefficient unit has a Swedish peer. Two of the four Finnish peers have no Swedish units in their referencing sets. Three peers, one each from Finland, the Netherlands and Norway, have the maximal number of inefficient Swedish units in their Referencing sets. Actually the Finnish and Norwegian

Table 3. The degree of localness index (5).
Country partitioning (4) of Referencing sets (3)

	Denmark		**Finland**				**the Netherlands**			**Norway**		**Sweden**	
	1009	1023	2014	2016	2026	2124	3005	3010	3017	4192	4462	5022	5047
Denmark	10	21	13	4	4	8	5	12	4	6	0	1	0
Finland	8	3	15	3	13	2	2	12	0	9	0	2	0
Netherlands	6	11	6	0	7	0	2	6	1	7	0	0	0
Norway	2	3	12	4	3	1	0	5	0	15	0	8	0
Sweden	1	1	33	0	0	8	0	28	2	34	0	30	6
Total count	27	39	79	11	27	19	9	63	7	71	0	41	6
Localness index	0.37	0.54	0.19	0.27	0.48	0.11	0.22	0.10	0.14	0.21	-	0.73	1.00

[8] The maximal number for each inefficient unit is five peers, since there are six constraints in (1) and the solution for the efficiency score is always positive.

peers are referencing more Swedish units than the Swedish peers themselves, and the Dutch peer just a few Swedish units less than the Swedish peer with the highest number of Swedish inefficient units in its Referencing set.

We interpret the obtained values of the Degree of peer localness index as empirical support for the importance of international benchmarking. Furthermore, the frontier seems to be well supported by the data in the sense that there is only one self-evaluator among the peers (Norwegian peer 4462).

4.4. Cross-country peer patterns

While the Degree of localness index is peer-specific, there may also be a need for a description of how countries are interconnected. The results for the Cross-country peer importance index, ρ_{qr}^{s}, defined in (6), are set out in Table 4 (a-c).

As explained in Section 2 the index is based on combining the numbers in Table 3 on the occurrence of inefficient units in the referencing sets, and the weights, λ_{ip}, which are part of the solution to model (1).[9] The origin of the inefficient units is given by the rows, and the columns give the origin of peers. The interpretation of a cell number, e.g. 17.5 in the second cell in the first row of Panel a, is the relative share of the weighted input saving (in percent) of replacement value of capital of inefficient Danish units referenced by Finnish peers. If we look at the most influential peer country for the inefficient units we see that for two of the three inputs Dutch peers are more important than Danish ones for inefficient units in Denmark, while for Dutch inefficient peers Danish peers are the most important for one input and Dutch peers for two inputs. For Finnish inefficient units Finnish peers are the most important by a large margin, and this national influence is also the case for Norway. For Swedish inefficient units Finnish peers are the most important for all inputs, and then the Dutch peers, the Swedish peers coming third.

[9] The numbers are reported in Edvardsen and Førsund (2002).

Table 4.Cross-country peer importance index (6) in percent.
Panel a. Replacement value of capital

	Denmark	Finland	the Netherlands	Norway	Sweden
Denmark	39.8	17.5	37.8	0.3	4.7
Finland	2.1	72.4	20.3	0.7	4.5
Netherlands	45.4	10.9	40.6	3.1	0.0
Norway	15.5	23.7	5.7	43.1	12.0
Sweden	0.2	46.5	26.1	4.9	22.3

Panel b. Total operating and maintenance costs

	Denmark	Finland	the Netherlands	Norway	Sweden
Denmark	34.6	15.2	40.1	0.4	9.8
Finland	3.6	57.5	32.2	1.2	5.4
Netherlands	38.4	14.0	45.2	2.4	0.0
Norway	10.2	22.9	8.1	40.6	18.2
Sweden	0.6	36.7	33.0	4.5	25.3

Panel c. Energy loss

	Denmark	Finland	the Netherlands	Norway	Sweden
Denmark	36.7	16.1	38.4	0.3	8.4
Finland	3.1	60.1	31.3	1.2	4.3
Netherlands	38.3	14.9	44.4	2.5	0.0
Norway	9.1	22.6	8.7	44.0	15.5
Sweden	0.4	35.4	30.3	4.4	29.5

Inspecting the peer groups we see that the Danish peers are more important for Dutch inefficient units than for Danish ones, the latter coming second for all inputs. The Finnish peers are most important for Finnish inefficient units and then for Swedish units for all inputs. The Dutch peers are most important for Dutch units, and then come Danish inefficient units. Norway and Sweden are most strongly connected, with Norwegian peers having Swedish inefficient units as the second most important group of inefficient units after its own. Swedish peers also have Norwegian inefficient units as the second most important group after its own inefficient units.

The location of small and high values of the index values shows the pattern of cross-country connections. The Norwegian peers have a very low importance for inefficient units in all other countries than Norway itself and Sweden. As seen also from Table 3 there is no connection between Swedish peers and Dutch inefficient units. The impact on Danish and Finnish peers is small compared with the impact on Norwegian units and its own inefficient units. The connections between Denmark and the Netherlands work symmetrically both ways, while Finnish peers influence Swedish inefficient units much more than Swedish peers are of importance for Finnish inefficient units, and Dutch peers are more important for Finnish inefficient units than Finnish peers are for Dutch inefficient units.

4.5. Local versus common technology

We have investigated the possibility of operating with individual country technology by running the DEA model for the three output- and three input variables. However, we may have a problem of dimensionality with Denmark, Finland, the Netherlands and Norway, since this sample includes 24, 25, 16 and 14 units respectively. The ad hoc rule (Cooper et al., 2000) that there are dimensionality problems if the number of dimensions multiplied with three is higher than the number of observations, apply to the Netherlands and Norway. A run of country specific technologies is presented together with the common frontier in Figure 2. The ordering of units within the countries from the common technology run is kept, and the scores for country specific technologies shown by the step curve above the bars are ordered identically. As expected the number of efficient units in the Netherlands and Norway increase drastically, and also for Denmark. The individual changes for the units can be large, illustrating the dimensionality problem for all countries except Sweden. The distribution for Sweden with 42 observations is much more stable, and we see a more or less parallel shift upwards of the whole distribution. Of the 11 units being efficient within the local frontier, only two remain so within the common frontier, and only one peer has other countries' inefficient units in its referencing set. Other countries' peers setting a higher standard for Swedish units cause the downward shift the efficiency distribution. The importance of exposing national peers to international benchmarking is clearly demonstrated.

4.6. Productivity comparisons of countries

Table 5 shows the ratios of the geometric average of the efficiency scores for each country relative to all other countries and also to the total geometric mean (cf. Equations (9) and (10)). Finland seems to be the most productive country within the common technology, having a bilateral index value compared with all the other countries higher than one. Sweden comes closest, while Norway and the Netherlands are on about the same level, and Denmark is the least productive country. Starting with Denmark; Finland and Sweden are the most productive countries relative to it, while the Netherlands and Norway are ahead by 4 to 6 percentage points. Norway's performance is closest to the Netherlands', lagging it by about 1 percentage point. It is interesting to note, in view of the special situation for Sweden revealed earlier, that Sweden, after all, on average, is in front of all countries with the exception of Finland. We can use the performance against the total sample average as a final ranking. The last row shows that the ranking has Finland at the top, then Sweden, the Netherlands, Norway and Denmark, the two first countries being in front of the total (geometric) average and the other three behind. The use of the Malmquist indices to rank countries corresponds closely to the form of the country efficiency distributions discussed above in connection with Figure 2, where it was pointed out that Finland and Sweden had the most even distributions and the highest share of units above the total sample mean.

5. Conclusions

When doing international benchmarking for the same type of production activity in several countries, applying a common frontier technology seems to yield the most satisfactory environment for identifying multinational peers and assessing the extent of inefficiency. In our exercise for a sample of large electricity distribution utilities from Denmark, Finland Norway, Sweden and the Netherlands it is remarkable that peers come from all countries. The importance of exposing national units, and especially units that would have been peers within a national technology, to international benchmarking is clearly demonstrated. The multinational setting has called for the development of new indices to capture the cross-

Table 5. Productivity comparisons of countries.
Malmquist productivity indices (9), (10)
calculated as ratios of geometric means

	Denmark	Finland	the Netherlands	Norway	Sweden
Denmark	1.00	1.16	1.06	1.04	1.12
Finland	0.86	1.00	0.91	0.90	0.97
Netherlands	0.95	1.10	1.00	0.99	1.06
Norway	0.96	1.11	1.01	1.00	1.08
Sweden	0.89	1.04	0.94	0.93	1.00
Average all units	0.92	1.07	0.97	0.96	1.03

country pattern of the nationality of peers and the nationality of units in their referencing sets. Bilateral Malmquist productivity comparisons can be performed between units of particular interest in addition to country origin, e.g. sorting by size, or location of utility (urban - rural), etc. We have focused on a single unit against the (geometric) average performance of all units, as well as bilateral comparisons of (geometric) averages of each country. Our results point to Finland as the most productive country within the common technology. This result reflects the more even distribution of the Finnish units and the high share of units above the total sample mean of efficiency scores.

The advantage of working with the DEA model is the richness of details available from the model solutions and the concrete connections to actual units. However, this may also be a problem because it is not always so easy to find explanations for specific findings such as why some units are efficient. The main practical purpose of the paper is to serve as a pilot study for the Nordic electricity regulators and the Dutch regulator as a start of a process of finding tools for regulation. The quality of the data and the acceptance of the model framework are of crucial importance for regulation since the units are regulated based on their individual performance as portrayed by the model results. There is at present some disagreement about the possibility of basing regulation of utilities on the approach use here (Shuttleworth (1999), Nillesen and Telling, 2001).

In order to improve upon the model approach as a benchmarking tool we would like to follow up the developments and results of the present study with the following research agenda:

i) Find explanations for the cross country peer and inefficient unit patterns revealed by the novel cross country peer importance indices

ii) Improve the comparability of data between countries by harmonizing definitions of variables and extending collection to cover environmental variables

iii) Define financial variables and collect data for cost efficiency exercises

iv) Investigate the scale properties by specifying variable returns to scale technology

v) Increase the number of cross section observations to cover all units within a country enabling country specific technologies also to be studied (if the total number of national units allows)

vi) Establish time series of cross sections enabling productivity developments to be studied

vii) Develop a more general transitive Malmquist index for the latter two cases

References

Andersen, P. and N. C. Petersen, 1993, A procedure for ranking efficient units in data envelopment analysis, Management Science 39, 1261-1264.

Berg, S. A., F. R. Førsund and E. S. Jansen, 1992, Malmquist indices of productivity growth during the deregulation of Norwegian banking, Scandinavian Journal of Economics 94, Supplement, 211-228.

Berg, S. A., F. R. Førsund, L. Hjalmarsson and M. Suominen, 1993, Banking efficiency in the Nordic countries, Journal of Banking and Finance 17, 371-388.

Caves, D.W., L.R. Christensen and E. Diewert, 1982a, The economic theory of index numbers and the measurement of input, output, and productivity, Econometrica 50, 1393-1414.

Caves, D. W., L.R. Christensen and W. E. Diewert, 1982b, Multilateral comparisons of output, input, and productivity using superlative index numbers, Economic Journal 92 (March), 73-86.

Charnes, A., W.W. Cooper and E. Rhodes, 1978, Measuring the efficiency of decision making units, European Journal of Operational Research 2(6), 429-444.

Cooper, W.W., L.M. Seiford and K. Tone, 2000. Data envelopment analysis. A comprehensive text with models, applications, references and DEA – solver software (Kluwer Academic Publishers, Boston/Dordrecht/London).

Edvardsen, D. F. and F. R. Førsund, 2002, International benchmarking of electricity distribution utilities, Working Paper No 08/02 ICER [http.//www.icer.it/docs/wp2002/ forsund08-02.pdf].

Farrell, M. J., 1957, The measurement of productive efficiency, Journal of the Royal Statistical Society, Series A, 120, III, 253-281.

Färe, R. and D. Primont, 1995, Multi-output production and duality: theory and applications (Kluwer Academic Publishers, Boston).

Färe, R., S. Grosskopf, B. Lindgren and P. Roos, 1994, Productivity developments in Swedish hospitals: a Malmquist output index approach, in: A. Charnes, W. W. Cooper, A.Y. Lewin and L. M. Seiford, eds., Data envelopment analysis: theory, methodology, and applications (Kluwer Academic Publishers, Boston/Dordrecht/London), 253-272.

Førsund, F. R., 1993, Productivity growth in Norwegian ferries, in: H. Fried, C. A. K. Lovell, and S. Schmidt, eds., The measurement of productive efficiency, techniques and applications (Oxford University press, Oxford), 352-373.

Førsund, F. R., 1999, On the contribution of Ragnar Frisch to production theory, Rivista Internazionale di Scienze Economiche e Commerciali (International Review of Economics and Business) XLVI (1), 1-34.

Førsund, F. R., 2002, On the circularity of the Malmquist productivity index, Working Paper No 29/02, ICER [http.//www.icer.it/docs/wp2002/forsund29-02.pdf].

Førsund, F. R. and S. A. C. Kittelsen, 1998, Productivity development of Norwegian electricity distribution utilities, Resource and Energy Economics 20(3), 207-224.

Hjalmarsson, L. and A. Veiderpass, 1992a, Efficiency and ownership in Swedish electricity retail distribution, Journal of Productivity Analysis 3, 7-23.

Hjalmarsson, L. and A.Veiderpass, 1992b, Productivity in Swedish electricity retail distribution, Scandinavian Journal of Economics 94, Supplement, 193-205.

Jamasb, T. and M. Pollitt, 2001, Benchmarking and regulation: international electricity experience, Utilities Policy 9(3), 107-130.

Johansen, L. and Å. Sørsveen, 1967, Notes on the measurement of real capital in relation to economic planning models, The Review of Income and Wealth, Series 13, 175-197.

Kittelsen, S. A. C., 1993, Stepwise DEA; choosing variables for measuring technical efficiency in Norwegian electricity distribution, Memorandum No. 6/1993. Department of Economics, University of Oslo.

Langset, T. og S. A. C. Kittelsen, 1997, Forsyningsareal og metodevalg ved beregning av effektivitet i elektrisitetsfordeling [Service area and choice of method when calculating efficiency in electricity distribution], Rapport 85. Stiftelsen for samfunns- og næringslivsforskning, Oslo.

Neuberg, L. G., 1977, Two issues in the municipal ownership of electric power distribution systems, Bell Journal of Economics 8(1), 303-323.

Nillesen, P. and J. Telling, 2001, Benchmarking distribution companies, EPRM Electricity March 2001, 10-12 [http.//www.icfconsulting.com/Publications/doc_files/ BenchmarkingDistributionCompanies.pdf].

Salvanes, K. G. and S. Tjøtta, 1994, Productivity differences in multiple output industries: an empirical application to electricity distribution, Journal of Productivity Analysis 5, 23-43.

Schaffnit, C., D. Rosen and J. C. Paradi, 1997, Best practice analysis of bank branches: an application of DEA in a large Canadian bank, European Journal of Operational Research 98, 269-289.

Shuttleworth, G., 1999, Energy regulation brief, National Economic Research Associates, n/e/r/a, London, 1-4 [http.//www.nera.com/wwt/newsletter_issues/4030.pdf].

Torgersen, A. M., F. R. Førsund and S. A. C. Kittelsen, 1996, Slack adjusted efficiency measures and ranking of efficient units, Journal of Productivity Analysis 7, 379-398.

Weiss, L.W., 1975, Antitrust in the electric power industry, in: A. Phillips, ed., Promoting competition in regulated markets (Brookings Institute, Washington, DC).

FAR OUT OR ALONE IN THE CROWD:

CLASSIFICATION OF SELF-EVALUATORS IN DEA[*]

by

Dag Fjeld Edvardsen

The Norwegian Building Research Institute,

Finn R. Førsund[†]

Department of Economics University of Oslo/

The Frisch Centre

and

Sverre A. C. Kittelsen

The Frisch Centre

Abstract: The units found strongly efficient in DEA studies on efficiency can be divided into self-evaluators and active peers, depending on whether the peers are referencing any inefficient units or not. The contribution of the paper is to develop a method for classifying self-evaluators based on the additive DEA model into interior and exterior ones. The exterior self-evaluators are efficient "by default"; there is no firm evidence from observations for the classification. These units should therefore not been regarded as efficient, and should be removed from the observations of efficiency scores when performing a two-stage analysis of explaining the distribution of the scores. The application to municipal nursing- and home care services of Norway shows significant effects of removing exterior self-evaluators from the data when doing a two-stage analysis.

Keywords: Self-evaluator, interior and exterior self-evaluator, DEA, efficiency, referencing zone, nursing homes

JEL classification: C44, C61, D24, I19, L32

[*] The paper is based on results from the NFR-financed Frisch Centre project "Better and Cheaper?" and written within the project "Efficiency analyses of the nursing and home care sector of Norway" at the Health Economics Research Programme at the University of Oslo (HERO) and the Frisch Centre.

[†] Corresponding author. Email: f.r.forsund@econ.uio.no, postal address: Department of Economics, University of Oslo, Box 1095, 0317 Blindern, Oslo, Norway.

1. Introduction

The calculation of efficiency scores for production units based on a non-parametric piecewise linear frontier production function, is well established within the last two decades. Originally introduced by Farrell (1957) the method was further developed in Charnes, Cooper and Rhodes (1978), where the term the *Data envelopment analysis (DEA) model* was coined. The efficient units span the frontier, but the classification of some of these units as efficient is not based on other observations being similar, but is due to the method. We are referring to units, which are classified as being *self-evaluators* in the literature a concept introduced by Charnes et al. (1985a). Self-evaluators may most naturally appear at the "edges" of the technology, but it is also possible that self-evaluators appear in the interior. It may be of importance to distinguish between those self-evaluators that are *exterior* and those that are *interior*. Finding the influence of some variables on the level of efficiency by running regressions of efficiency scores on a set of potential explanatory variables, is an approach often followed in actual investigations.[1] Using exterior self-evaluators with efficiency score of 1 may then distort the results, because to assign the value of 1 to these self-evaluators is arbitrary. Interior self-evaluators, on the other hand, may have peers that are fairly similar. They should therefore not necessarily be dropped when applying the two- stage approach.

The plan of the paper is to review the DEA models in Section 2 and define the new concepts of interior and exterior self-evaluators. In Section 3 the method for classifying the self-evaluators is introduced. Actual data are presented in Section 4 and the method for classifying self-evaluators is applied. The effect of removing exterior self-evaluators is shown. Section 5 concludes.

[1] The approach was originally introduced in Seitz (1967), inspired by Nerlove (1965), see Førsund and Sarafoglou (2002). Simar and Wilson (2003) review the approach and find it at fault in general due to serial correlation between the efficiency scores, and provides a new statistically sound procedure based on specifying explicitly the data generating process and bootstrapping to obtain confidence intervals**.**

2. Self-evaluators

DEA models

Consider a set, *J*, of production units transforming multiple inputs into multiple outputs. Let y_{mj} be an output $(m \in M, j \in J)$ and x_{nj} an input $(n \in N, j \in J)$. As the reference for the units in efficiency analyses we want to calculate a piecewise linear frontier based on observations, fitting as closely as possible and obeying some fundamental assumptions, like free disposal, and the technology set being convex and closed as usually entertained (Banker et al., 1984, Färe and Primont, 1995). This frontier can be found by solving the following LP problem, termed the *additive model* in the DEA literature (Charnes et al., 1985b):

$$
\begin{aligned}
&Max \quad \left\{ \sum_{m \in M} s_{mi}^{+} + \sum_{n \in N} s_{ni}^{-} \right\} \\
&s.t. \\
&\sum_{j \in J} \lambda_{ij} y_{mj} - y_{mi} - s_{mi}^{+} = 0 \quad , m \in M \\
&x_{ni} - \sum_{j \in J} \lambda_{ij} x_{nj} - s_{ni}^{-} = 0 \quad , \quad n \in N \\
&s_{mi}^{+} \quad , \quad s_{ni}^{-} \geq 0 \\
&\lambda_{ij} \geq 0 \\
&\sum_{j \in J} \lambda_{ij} = 1
\end{aligned}
\tag{1}
$$

The last equality constraint in (1) imposes variable returns to scale (VRS) on the frontier, while dropping this constraint imposes constant returns to scale (CRS). Our analysis will be valid for both scale assumptions. The frontier is found by maximising the sum of the slacks on the output constraints, s_{mi}^{+}, and input constraints, s_{ni}^{-}. The *strongly efficient* units (using the terminology of Charnes et al., 1985b) are identified by the sum of the slacks and therefore all the slack variables being zero. All weights, λ_{ij}, must be zero except the weight for itself that will be one (i.e. $\lambda_{ij} = 0 \ for \ i \neq j, \lambda_{ii} = 1$ if *i* is an efficient unit).[2] The efficient points will appear as vertex points on the frontier function surface, or corner points of facets. The sets of strongly efficient units, *P*, and the inefficient units, *I*, are:

[2] A strongly efficient unit, *i*, may end up being located exactly on a facet. We may then have multiple solutions for the weights, although the maximal sum of slacks is still zero. One of the solutions will be $\lambda_{ij} = 0$ for $j \neq i$, and $\lambda_{ii} = 1$.

$$P = \left\{ i \in J : \sum_{m \in M} s_{mi}^{+} + \sum_{n \in N} s_{ni}^{-} = 0 \right\}$$
$$I = \left\{ i \in J : \sum_{m \in M} s_{mi}^{+} + \sum_{n \in N} s_{ni}^{-} > 0 \right\} \ , \qquad (2)$$
$$P \cup I = J$$

So far we only have slacks as measures of inefficiency. If we want only one measure for each unit, and a measure that is independent of units of measurement, the Farrell (1957) measure of technical inefficiency is the natural choice. The standard DEA model on primal (enveloping) form, is set up as a problem of determining the Farrell technical efficiency score, E_{oi}, (o = 1,2), either in the input- (o = 1) or the output (o = 2) direction for an observation, *i*. The following LP model is formulated for each observation in the case of input-orientation:

$$\begin{aligned}
& E_{1i} \equiv Min \quad \theta_i \\
& s.t. \\
& \sum_{j \in P} \lambda_{ij} y_{mj} - y_{mi} \geq 0 \quad , m \in M \\
& \theta_i x_{ni} - \sum_{j \in P} \lambda_{ij} x_{nj} \geq 0 \quad , \ n \in N \qquad (3) \\
& \lambda_{ij} \geq 0 \\
& \sum_{j \in P} \lambda_{ij} = 1
\end{aligned}$$

In the case of output orientation we have the following LP program:

$$\begin{aligned}
& 1 / E_{2i} \equiv Max \quad \phi_i \\
& s.t. \\
& \phi_i y_{mi} - \sum_{j \in P} \lambda_{ij} y_{mj} \leq 0 \quad , m \in M \\
& \sum_{j \in P} \lambda_{ij} x_{nj} - x_{ni} \leq 0 \quad , \ n \in N \qquad (4) \\
& \lambda_{ij} \geq 0 \\
& \sum_{j \in P} \lambda_{ij} = 1
\end{aligned}$$

For notational ease the same symbols have been used for weights in (1), (3) and (4). The proportionality factor, θ_i or ϕ_i, and the weights, λ_{ij} , are the endogenous variables.

Adopting the notation #N and #M for the number of inputs and outputs respectively, the point

$$(\sum_{j \in P} \lambda_{ij} x_{1j}, .., \sum_{j \in P} \lambda_{ij} x_{\#Nj,} \sum_{j \in P} \lambda_{ij} y_{1j}, .., \sum_{j \in P} \lambda_{ij} y_{\#Mj}) \qquad (5)$$

is per construction on the frontier surface, and is defined as the *reference point* for unit *i*. If there are no slacks on the output- and input constraints in (3) or (4) then the reference point coincide with the radial projection point, using either θ_i or ϕ_i when adjusting an inefficient observation. These points will normally be interior points on facets (but may fall on border lines). With one or more slacks positive the reference point and the radial projection point differ. The reference points will again appear as vertex points on the frontier function surface, or corner points of facets.

It is well known that the radial Farrell efficiency measure E_{oi} may be one, but that the unit may still improve its performance by either using less inputs or producing more outputs. All units with a radial efficiency score of one are by definition located on the frontier, but it is only for the *strongly efficient* units that the reference points coincide with the observation. A unit may have E_{oi} =1, but one or more of the constraints in (3) or (4) being non-binding (i.e. one or more slacks positive and zero shadow prices on the constraints in question).

Although the model (3) or (4) can be solved directly by letting the index *j* run over all observations in *J*, a two-stage procedure of solving (1) first is often followed. By using the information on strongly efficient units when solving (3) or (4), the LP computations are done more efficiently, and one will only identify reference points by (5) that are in the strongly efficient subset of the frontier.

In the context of the DEA models (3) and (4), the strongly efficient units are termed *peers*. For each inefficient unit, *i*, a *Peer group set*, P_i, (Cooper, Seiford and Tone, 2000) may be formed:

$$P_i = \{p \in P : \lambda_{ip} > 0\}, i \in I \tag{6}$$

where λ_{ip} are the solution values of the weights in either (3) or (4) depending on the orientation in question. If the Peer group sets are empty, then all the units are efficient. The solutions to (1), (3) or (4) do not identify facets systematically, but by using (6) we can identify the corner points of facets where one or more radial projection points of inefficient units are located.

It will also turn out useful to look at the group of inefficient units referenced by a peer. Such a set is defined for each peer, *p*, as the *Referencing set* in Edvardsen and Førsund (2001) with reference to the solutions of (3) or (4):

$$I_p = \left\{ i \in I : \lambda_{ip} > 0 \ \right\},\ \ p \in P \tag{7}$$

The self evaluators

The Referencing set (7) may be empty, in which case the unit is called a self-evaluator:

Definition 1: A peer $p \in P$*, where the set P is defined in (2), is a self-evaluator if* $I_p = \emptyset$*, where* I_p *is defined in (7)*[3].

The set of peers may thus be partitioned into a set of self-evaluators, P^S, and a set, P^A, of *active peers,* i.e. peers with non-empty referencing sets:

$$\begin{aligned} &P^S = \left\{ p \in P : I_p = \emptyset \right\} \\ &P^A = \left\{ p \in P : I_p \neq \emptyset \right\} \\ &P^S \cup P^A = P \end{aligned} \tag{8}$$

The self-evaluators are vertex points of facets without any reference points defined as the radial projection points of inefficient observations located on these facets. The LP solutions to (3) or (4) do not give us any information as to which efficient units constitute the vertex points of such a facet without reference points. An efficient unit may be a vertex point for many facets. Our definition of a self-evaluator implies that there are no reference points on any of its facets.

3. The determination of type of self-evaluator

There are two possibilities as to the location of facets formed by self-evaluators on the frontier surface. Such facets may be part of the extreme areas of the frontier, i.e. facets closest to the axes in the case of CRS, or facets, in the case of VRS, also furthest away from the origin or closest to the origin (the VRS frontier will in general not contain the origin). In the case of

[3] An alternate definition could be in terms of the reference shares defined in Torgersen, Førsund and Kittelsen (1996), where a self-evaluator has a reference share of zero.

CRS only mixes of inputs or outputs may be extreme, while in the case of VRS we in addition have the scale dimension. Such self-evaluators will be termed *exterior* self-evaluators. In the case of CRS, facets without any reference points may also be found in the interior of the frontier surface with respect to mixes, while for VRS interior also means interior regarding scale. Such self-evaluators will be termed *interior* self-evaluators.

Figure 1 shows the two different cases in the simplest case of two dimensions. The observations represented by points *A, B, C, D, F* and *G* are efficient, while O_1 is inefficient. The radial reference or projection point for unit O_1 is *a* in the case of input orientation. The reference point (5) in this simple case coincides with the peer *A*. Considering output-orientation the peers are *D* and *F*, and the reference point is *d*. To illustrate the referencing set of a peer, the shaded area in Figure 1 shows the *referencing zone* for the efficient unit *D* in the case of output orientation. All the inefficient units being in unit *D*'s referencing set must be located here (such inefficient units may also appear in referencing sets of other peers; here unit *F*'s). If the referencing zone is empty then the peer is a self-evaluator. Removal of such a self-evaluator will not change the efficiency scores for any other units. We would expect the self-evaluators to be extreme points in one or more of the mix or scale dimensions, but if the referencing zone is narrow a self-evaluator may also be centrally placed within the set of observations. A narrow zone means that other peers are close to the self-evaluator.

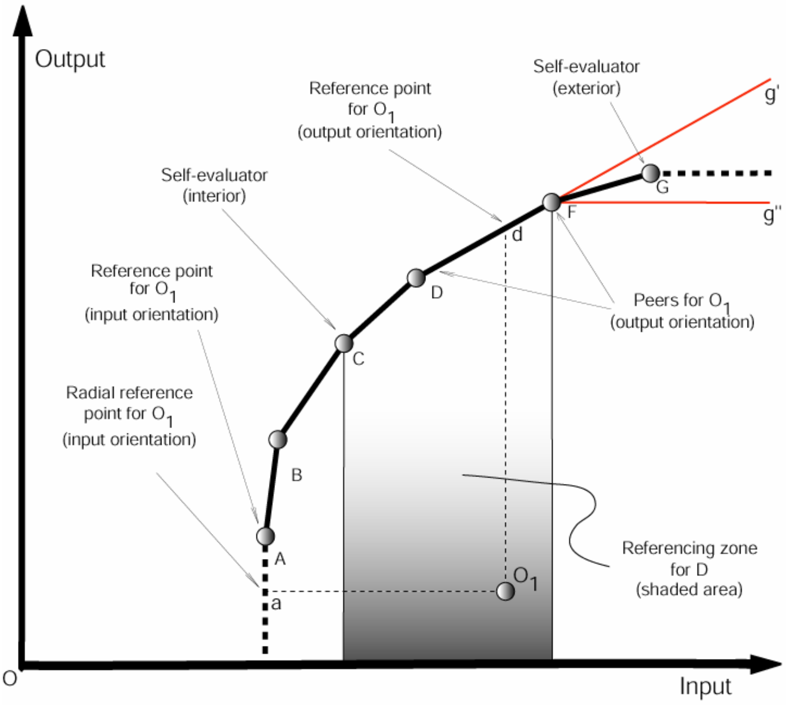
Output
Self-evaluator (exterior)
Reference point for O1 (output orientation)
g'
Self-evaluator (interior)
G
F
g''
d
Reference point for O1 (input orientation)
D
Peers for O1 (output orientation)
C
Radial reference point for O1 (input orientation)
B
A
a
O1
Referencing zone for D (shaded area)
O
Input

Figure 1: DEA and the two types of self evaluators

Notice that the classification as a self-evaluator is dependent of the orientation of the efficiency measure. Considering output orientation we have that both *B* and *C* are interior self-evaluators, while *A* and *G* are exterior self-evaluators. Considering input orientation we have that *B*, *C*, *D* and *F* are interior self-evaluators, while *G* is an exterior one. In both cases the unit G could have been observed anywhere between the line g' (the continuation of the line DF) and the line g'' (referenced by F), without any unit changing its estimated efficiency or its status as peer. The efficiency score of 1 assigned to unit G therefore contains little information. In e.g. the output oriented case we see that there is a considerable scope for output variation for a given input yielding the efficiency score of 1.

Our purpose is to develop a method for classification into exterior or interior self-evaluators using only the standard DEA format.

Enveloping from below

The production set is by construction convex. If all inefficient units are removed from the data set, and a new run is done with only the efficient units, we will find the exterior peers by reversing the enveloping of the data from "above" to be from "below". All that needs to be done is to reverse the inequalities in the LP program (1) by *adding* the slack variables instead of subtracting:

$$\begin{aligned}
&Max \quad \left\{ \sum_{m\in M} s_{mi}^{+} + \sum_{n\in N} s_{ni}^{-} \right\} \quad (i \in P) \\
&s.t. \\
&\sum_{j\in P} \lambda_{ij} y_{mj} - y_{mi} + s_{mi}^{+} = 0 \quad ,m \in M \\
&x_{ni} - \sum_{j\in P} \lambda_{ij} x_{nj} + s_{ni}^{-} = 0 \quad , \ n \in N \\
&s_{mi}^{+} \ , \quad s_{ni}^{-} \geq 0 \quad ,m \in M \quad , \ n \in N \\
&\lambda_{ij} \geq 0 \\
&\sum_{j\in P} \lambda_{ij} = 1
\end{aligned} \tag{9}$$

Notice that we are only considering observations belonging to the set of strongly efficient units *P* determined by solving (1). This envelopment of the data is by construction concave.

The units that turn out as “efficient” in solving (9), in the sense that all slacks are zero, must be units belonging to the exterior facets in the solution to the original model (1). We will use this result to define exterior and interior strongly efficient units:

Definition 2: A strongly efficient unit belonging to the set P defined by (2) is exterior if it belongs to the set P^E:

$$P^E = \left\{ p \in P : \sum_{m \in M} s_{mp}^{+} + \sum_{n \in N} s_{np}^{-} = 0 \right\} \qquad (10)$$

where the slack variables, s_{mp}^{+}, s_{np}^{-} , are solutions to the problem (9).

A strongly efficient unit belonging to the set P defined by (2) is interior if it belongs to the set

$$P^I = \left\{ p \in P : \sum_{m \in M} s_{mp}^{+} + \sum_{n \in N} s_{np}^{-} > 0 \right\} \; (P^E \cup P^I = P) \qquad (11)$$

where the set P^E is defined in (10)[4].

To determine the nature of a self-evaluator an orientation for the calculation of the Farrell efficiency measures has to be chosen, i.e. either input- or output orientation. The following definition can then be made as to the classification of self-evaluators:

Definition 3: Consider a peer $p \in P$, where the set P is defined in (2), that is a self-evaluator, $p \in P^S$, where the set P^S is defined in (8) and found by running either the input-oriented program (3), or the output-oriented program (4). If $p \in P^E$, where the set P^E is defined in (10), then p is an exterior self-evaluator. If $p \notin P^E$ then p is an interior self-evaluator:

$$\begin{aligned} &P^{SE} = P^S \cap P^E \\ &P^{SI} = P^S \cap P^I \qquad (12) \\ &(P^{SE} \cup P^{SI} = P^S) \end{aligned}$$

where P^{SE} and P^{SI} are the sets of the exterior and interior self-evaluators respectively.

Illustrating the approach using Figure 1, we have that the new “from below frontier” will be the line from *A* to *G*, thus these units are the only ones on the “from below frontier” and

[4] Note that in the special case where two units have identical input-output vectors, both could be classified as exterior by this criterion, but would not have unique intensity weights in (9). In this situation, which is likely to be rare in empirical applications, it seems natural to classify the units as interior.

therefore exterior points in P^E. This classification is independent of orientation, and they are both being located on exterior facets in the original problem (1). In the case of output orientation, the self-evaluators *B* and *C*, according to the solution to problem (4), will not appear on the new frontier, and they are therefore interior according to Definition 3. The self-evaluators *A* and *G* appear on the new frontier and are therefore exterior. In the case of input orientation solving problem (4) gives *B, C, D, F* and *G* as self-evaluators, and we have that *B, C, D, and F* are interior self-evaluators and *G* an exterior one. While *A* is an exterior peer in input orientation, it is not a self-evaluator.

Figure 2 provides another illustration. In a two-dimensional input space an isoquant is shown in the efficient units *A, B, C* and *D*. Consider input orientation and CRS. Assuming inefficient units are only located northeast of the isoquant segment *AB* in the cone delimited

by the rays going through the points *A* and *B*, we have that *C* is an interior self-evaluator, and *D* is an exterior self-evaluator. Running the "reverse" program (9) we will envelope the four peers from "behind" by the broken line from *A* to *D*. We then know that units *A* and *D* are exterior, and using the information from running the DEA model (1) we then have that unit *C* is an interior self evaluator, and unit *D* an exterior one.

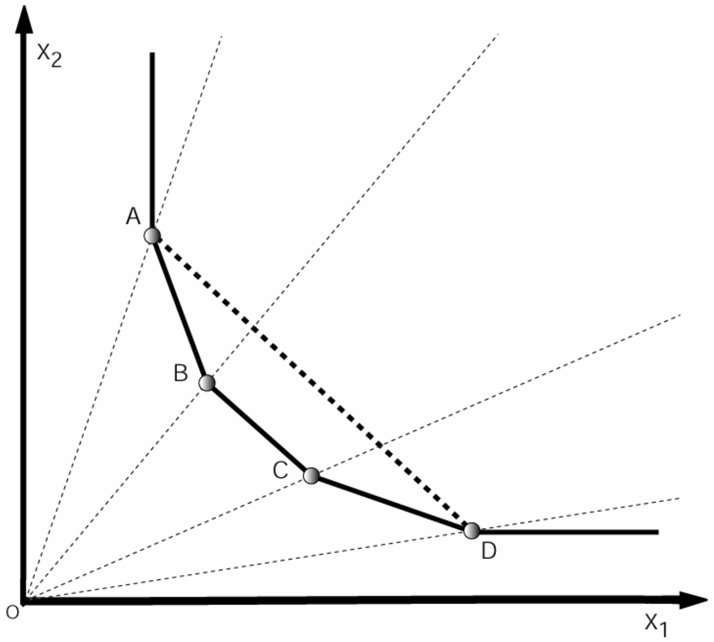

Figure 2. Determining the type of self-evaluator

It may also be of interest to classify the active peers according to the type exterior and interior. Building on definition 3 we have.

Definition 4. The active peers defined in (8) belong to the subsets P^{AE} and P^{AI}:

$$\begin{aligned} &P^{AE} = P^{A} \cap P^{E} \\ &P^{AI} = P^{A} \cap P^{I} \qquad (13) \\ &(P^{AE} \cup P^{AI} = P^{A}) \end{aligned}$$

where P^{E} and P^{I} are defined in (10) and (11) respectively.

The program (9) is not the standard DEA additive formulation, since the sign of the slacks in the restrictions on inputs and outputs have been changed. However, by negating these equalities, (9) can be rewritten as:

$$\begin{aligned} &Max \quad \left\{ \sum_{m\in M} s_{mi}^{-} + \sum_{n\in N} s_{ni}^{+} \right\} \quad (i \in P) \\ &s.t. \\ &\sum_{j\in P} \lambda_{ij} x_{nj} - x_{ni} - s_{ni}^{+} = 0 \quad , n \in N \\ &y_{mi} - \sum_{j\in P} \lambda_{ij} y_{mj} - s_{mi}^{-} = 0 \quad , \quad m \in M \qquad (14) \\ &s_{mi}^{+} \quad , \quad s_{ni}^{-} \geq 0 \quad , m \in M \quad , \quad n \in N \\ &\lambda_{ij} \geq 0 \\ &\sum_{j\in P} \lambda_{ij} = 1 \end{aligned}$$

Comparing (1) and (14) we see that these are identical except that inputs and outputs are exchanged. Since existing DEA software often will solve the additive model (1), we may as well for convenience find the set of exterior self-evaluators P^{SE} by exchanging inputs and outputs and running (14) on the strongly efficient units, rather than running (9) on these units.

4. An empirical application

The data

We will apply the method for determining interior and exterior self-evaluators on a cross section data set of the nursing and home care sector of Norwegian municipalities. The data is found in Edvardsen et al. (2000). The primary data source is the official yearly statistics for municipal activities published by Statistics Norway. Resource usage is measured by financial data and number of man-years of different categories. Production data contains mainly the number of clients dealt with by institutionalised nursing, home based nursing, and practical assistance. Quality information is lacking, but the clients are split into some age groups that may be of significance for resource use. In cooperation with representatives form the municipalities and the ministries of Finance, Municipal and regional affairs, and Social and health affairs we have chosen to split the clients on two major age groups, 0-66 and above 66 (67+), and use institutions and home care as separate outputs. Within institutions there are also a number of short-stay clients, either coming on a day care basis or on limited stay of convalescence. These usually require fewer resources than the permanent clients. As indicators of quality of institutions we have information of number of single person rooms and on clients staying in closed wards. The separation is regarded both as a quality factor for the clients taken care of (demented cases), and for the other clients. In home-based care mentally disabled may be quite resource demanding. They may also be found in the 0-66 age group within institutions. There is no information on how long time a home visit may last or how often it is received. Such information would obviously have given us some quality indicators. We also run the risk of municipalities cutting down on both length and number of visits showing the same number of clients receiving a more generous support in other municipalities.

To ensure that the data quality was good enough we entered a phase of quality control. We strongly feel that one should not automatically remove outliers, but if possible contact the municipality in question and ask if the data is correct. This is especially important if the methodology is frontier based (such as DEA) because the units defining the frontier are outliers by definition. This led to many changes in the dataset and required quit a lot of work, but as a result we could be much more confident in the quality of the data (see Aas (2000) for details).

Table 1: Primary variables used in the DEA model, cross-section 1997 of 469 municipalities.

		Average	**Standard deviation**	**Min**	**Max**
Inputs					
Trained Nurses	x_1	31.1	41.4	1.5	410.4
Other Employees	x_2	137.4	169.4	5.3	1821.5
Other expenses	x_3	9066.2	13449.5	190.0	108990.0
Outputs: No. of Clients					
Institutions, age 0-66	y_1	3.4	4.9	0.0	50.0
Institutions, age 67+	y_2	87.7	108.6	0.0	1024.0
Short-term stay	y_3	113.8	163.3	0.0	1614.0
Closed wards	y_4	11.8	19.3	0.0	195.0
Single person room	y_5	65.7	82.2	0.0	747.0
Mentally disabled	y_6	48.7	79.5	0.0	857.0
Practical assistance, 0-66	y_7	51.3	66.3	0.0	597.0
Practical assistance, 67+	y_8	212.7	272.4	1.0	2190.0
Home based nursing, 0-66	y_9	34.1	45.3	0.0	407.0
Home based nursing, 67+	y_{10}	125.8	153.3	1.0	1480.0

Table 1 shows descriptive statistics for the variables used in the DEA model. The first three rows measure the inputs in the model. *Trained nurses* and *Other Employees* shows us that about 18% of the employees (measured in man-years) are trained nurses. *Other expenses* are measured in 1000 NOK (Norwegian currency). The last 9 rows in table 1 measure the outputs. *Institution, age 0-66* and *Institutins, age 67+* are the number of institutionalized clients in the age groups 0-66 and above 67 respectively. *Short-term stay* shows how many visits the institutions in the municipality have gotten from clients who are not residents, while *Closed ward* shows how many of the residents are in a special ward for dementia clients. *Mentally disabled* shows how many of the clients are mentally disabled (almost all of these clients get home care). *Practical assistance, 0-66* and *Practical assistance, 67+* counts how many clients get practical assistance (such as cleaning and making food) in the indicated age groups, while *Home based nursing, 0-66* and *Home based nursing, 67+* count the same for clients getting nursing services in their own homes.

The Farrell output-oriented efficiency scores

Figure 3 shows E_2 (output-increasing efficiency assuming variable returns to scale). Each bar in the diagram represents one of the 469 municipalities, sorted by increasing efficiency. The

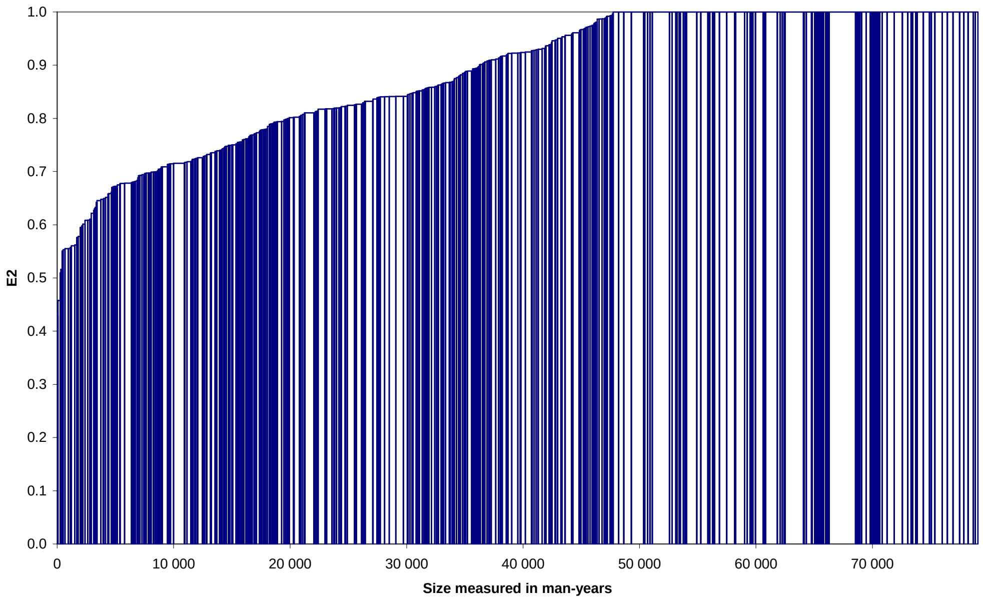

Figure 3: Sorted output-oriented efficiency scores

height of the bars represents the efficiency of the DMU, while the width of the bar shows the size measured by man-years (sum of trained nurses and other employees). Both large and small DMUs can be found in all parts of the diagram, with the exception that no large municipalities are located in the (very inefficient) leftmost part of the diagram. The average efficiency is 86 percent, while the efficiency of the average unit is 67 percent.

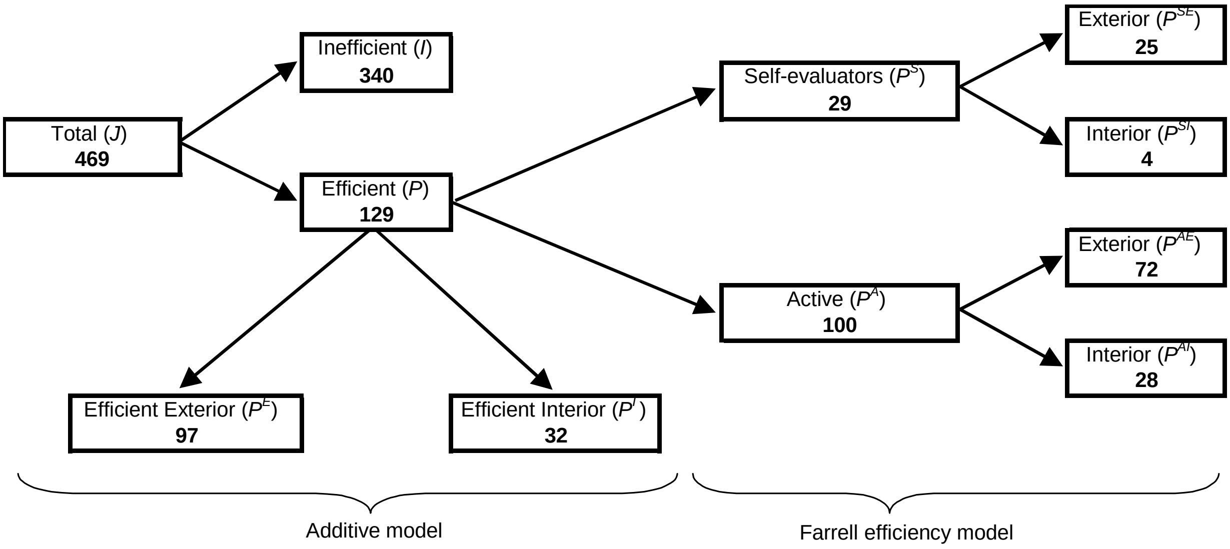

Figure 4: The taxonomy of units in DEA efficiency analyses

An overview of the taxonomy developed in Sections 2 and 3 for classification of units is given in Figure 4, together with the actual decomposition for the data set at hand. In view of the relatively large number of observations it may be surprising that as many as 28 percent of the units are efficient. This may be due to the unusually high number of dimensions, 13 variables in all. Since the efficient units span out the frontier technology it is to be expected that the number of exterior ones is higher than the interior ones, 75 and 25 percent respectively. Turning to the Farrell efficiency model (4) the self-evaluators represent 23 percent of the efficient units. As expected the relative share of exterior peers is larger in the group of self-evaluators than in the group of active peers, 86 versus 72 percent. Among the active peers that share of interior units is higher, 28 percent. This distribution is of importance for the empirical support of the frontier and the associated efficiency distribution.

Table 2. Relative size of interior- and exterior self-evaluators measured as percentage deviation from the sample average

Municipality number and name	Inputs			Outputs									
	Trained nurses	Other employees	Other expenses	Inst 0-66	Inst 67+	Short time stay	Closed ward	Single person room	Mentally disabled	PA 0-66	PA 67+	HS 0-66	HS 67+
Interior self-evaluators													
425 Åsnes	20	3	-1	-12	68	15	61	25	7	-2	27	6	54
616 Nes	-56	-63	-58	-41	-44	-38	-49	-25	-71	-40	-36	-38	-41
807 Notodden	14	36	22	17	83	167	53	106	15	-34	42	17	84
1567 Rindal	-51	-68	-64	-71	-41	-32	-41	-45	-84	-73	-54	-82	-39
Exterior self-evaluators													
101 Halden	126	194	62	17	32	-32	-100	14	198	167	319	193	402
213 Ski	63	43	66	311	-12	-38	69	11	54	257	58	-30	-3
217 Oppegård	92	6	29	76	-9	22	-100	28	42	15	43	390	184
219 Bærum	929	736	1102	825	748	551	1112	1038	781	659	538	577	415
430 Stor-Elvdal	-68	-48	-49	-100	-52	-49	-7	-50	-77	-67	-37	26	-42
615 Flå	-90	-73	-73	-100	-71	-77	-100	-70	-90	-84	-81	-97	-77
632 Rollag	-81	-72	-74	-71	-69	-89	-100	-63	-65	-80	-80	-24	-81
709 Larvik	345	296	205	517	283	303	273	241	393	356	390	170	280
806 Skien	353	356	346	252	297	93	19	304	235	349	551	322	562
904 Grimstad	92	21	-2	-12	27	-23	188	-7	54	159	44	225	9
941 Bykle	-86	-76	-77	-100	-73	-82	-58	-63	-98	-94	-96	-91	-94
1144 Kvitsøy	-85	-95	-94	-100	-95	-89	-100	-91	-100	-94	-94	-94	-90
1222 Fitjar	-61	-64	-73	-71	-70	-58	-32	-27	-77	-75	-73	-77	-79
1411 Gulen	-68	-49	-57	-41	-24	-84	-100	-57	-75	-84	-59	-65	-45
1612 Hemne	-51	-57	-58	-71	-65	-16	-7	-82	-88	-49	-41	38	7
1632 Roan	-88	-82	-86	-100	-82	-77	-100	-71	-84	-94	-79	-91	-74
1702 Steinkjer	201	116	27	135	85	87	358	68	60	157	172	126	115
1714 Stjørdal	128	96	29	487	97	99	36	128	153	126	60	3	18
1723 Mosvik	-72	-87	-89	-100	-76	-82	-100	-76	-77	-88	-85	-88	-75
1839 Beiarn	-78	-75	-74	-41	-75	-75	-49	-62	-90	-90	-71	-77	-65
1868 Øksnes	-48	-25	-48	-12	-37	-57	-41	-19	-42	-40	-56	143	-26
1920 Lavangen	-87	-82	-83	-71	-77	-82	-66	-73	-79	-96	-85	-97	-84
3001 Bygdøy-Frogner	41	10	150	-100	-37	-100	-100	-16	-32	15	194	99	188
3003 St.Hanshaugen-Ullevål	262	271	695	47	392	167	222	441	33	37	347	-9	231
3004 Sagene-Torshov	241	339	808	311	305	338	171	252	23	372	554	176	351

Far out or alone in the crowd?

In Table 2 the relative distance from the average unit is illustrated by measuring each of its variables against the average for the sample (J). The first four units are the interior self-evaluators in P^{SI} . These units are on both sides of the average, and one of the four units is quite close to the sample average. No unit is close to either the small or large exterior units. It seems appropriate to use the expression “alone in the crowd.”

The 28 exterior self-evaluators are distributed with half above and half below the sample average. One unit has maximal sample values for two of the variables. There are several output variables with zero as the lower limit. The variable “Institution, 0-66” has seven exterior units with the minimum value of zero; while for “Closed ward” there are eight exterior units with the minimum value of zero. So given that “far out “ means both small and large units the exterior units deserve well this classification. The influence of extreme mixes may also be investigated, but due to all the possible comparison we leave this exercise out.

The idea behind the two-stage approach is based on the distinction between pure inputs and outputs on the one hand, and environmental variables on the other. By the assumptions of the DEA method, the input-output vectors must belong to a deterministic technology set bounded by the frontier. However, environmental variables may be relevant for the performance of the units, but their influence may be regarded to be of a stochastic nature that is most appropriately revealed by studying the statistical association between some measure of performance and the environmental variables. Since the crucial point of being concerned with environmental variables is that there must be some influence on either the discretionary inputs or outputs of the environmental variables there is a good case for advocating a single stage approach and incorporating all relevant variables in one single model. One reason for treating environmental variables differently than standard outputs and inputs is that the way the variables interact with the standard production variables may be difficult to model. It may, for example, not be so clear-cut whether the variable is an input or an output.

The formulation of the second stage is to establish an association between the efficiency score and the environmental variables, z_k:

$$E_{oi} = f(z_1,..,z_K) + \varepsilon_i \ , i \in J, o = 1,2 \qquad (15)$$

where ε_i is a random variable. There have been several approaches to estimating (15). The first approach was to specify *f*(.) as a linear function and apply OLS (Seitz 1967, 1971). But there are two special features of the model (15). By definition the efficiency scores are restricted to be between zero and one,

$$0 \le E_{oi} = f(z_1,..,z_K) + \varepsilon_i \le 1 \;, i \in J \;, o = 1,2 \qquad (16)$$

and using the DEA model (3) or (4) to generate the efficiency scores usually leads to a concentration of the values 1. As shown in Figure 4 we have 28 percent of the efficiency scores at the upper limit of one. This has lead researchers to apply a censored regression like the Tobit model or truncated regressions. These approaches are strongly criticized in Simar and Wilson (2003). The fundamental point is made that the efficiency scores in (15) are estimates of the unknown efficiencies, and that these scores are serially correlated. Therefore, neither applying a Tobit or a truncated regression will solve this problem. A sequence of bootstrapping techniques is proposed that will yield proper confidence intervals of the parameters of *f*(.).

Table 3: Stage 2 regression results applying OLS to a linear model

	All units included*)		Excluding exterior Self-evaluators	
	R2	0.1737	R2	0.2082
Variable	Coeff.	p-value	Coeff.	p-value
Climate indicator	-0.007	0.035	-0.006	0.056
Share of private institutions	0.098	0.019	0.099	0.015
Free disposable income, 1996	-0.020	0.054	-0.028	0.010
Share of users in home care	-1.019	0.000	-1.089	0.000
Share in home care of age group 0-66	1.823	0.170	1.437	0.282
Share in home care of age group 67-79	0.574	0.006	0.665	0.001
Share in home care of age group 80-89	0.270	0.004	0.261	0.004
Share in home care of age group 90+	0.100	0.019	0.109	0.011
Share in inst. care of age group 0-66	24.785	0.019	27.615	0.009
Share in inst. care of age group 97-79	-1.072	0.011	-0.926	0.026
Share in inst. care of age group 80-89	-0.101	0.524	-0.152	0.331
Share in inst. care of age group 90+	0.026	0.561	0.053	0.235
Constant term	1.527	0.000	1.562	0.000

*) Communities within the two major cities Bergen and Oslo are aggregated and one unit is removed from the data set

However, since the purpose of our paper is to demonstrate the importance of the role of exterior peers, the relation (15) is here interpreted just to represent an investigation of association and not to be a causality model. Therefore, OLS is used to estimate a linear function (15). An advantage of OLS is that better diagnostics to characterize the covariations are available, like the multiple regression coefficient.

Table 3 shows the result of an OLS regression using a linear model in (15). The p-values are also given, although they should not be taken at face value due to the inherent statistical problems with the approach, as mentioned above. We perform regressions firstly with the complete data set, and secondly excluding the exterior self-evaluators.

The environmental variables represent background variables that experts have suggested may influence the efficiencies of municipalities. *Climate indicator* is a measure of the average temperature measured over the year in the municipality. It can also be seen as a proxy for amount of snow, altitude and distance from the coast. We note that removing the exterior self-evaluators changes both the regression coefficient and the p-value, indicating a weaker connection between efficiency scores and this variable.

Share of private institutions is measured by how large share of the total number of institutions are in the private sector (most often NGO's). It would be better to measure this by the number of clients, but such data was not available. Possible interpretations of a positive parameter estimate (and low p-values in both regression models) are that the municipalities own care providers get a learning effect from presence of private service providers, or that private presence reduces inefficiency because they increase the fear of privatization in the municipal nursing sector.

Free disposable Income, 1996 is measure of the relative wealth of the municipality (per inhabitant). It is calculated by finding the difference between the actual income in the municipality, and the "required expenses" in the municipality in other sectors than care for the elderly (i.e. schools, roads etc.). Required expenses are calculated on demographical variables and other factors exogenous to the municipality. (See Aaberge and Langørgen (2003) for the details behind the construction of this indicator.) Data for 1997 (the year all the other data is from) was also available, but we reasoned that the municipality's decision on how it want provide care for the elderly is more strongly based on income in the previous than in the current year. This has some statistical support in that the ’96 variable has larger explanatory power measured by R2 of the model and T-value of the parameter estimate. The p-value for the parameter estimate for this variable improves when the exterior self-evaluators are removed from the regression model. One possible explanation of the negative parameter estimate is that a “rich” municipality might use the extra resources on higher quality (not picked up by the DEA model) and/or allowing inefficiency in production of services.

Share of users of home care is a measure of the size of the share of home care clients in relation to all the clients getting nursing services. This coefficient has a negative parameter estimate. This is an indication that the technical efficiency tends to be lower when a larger part of the municipality’s clients is in home care. This is interesting, because it is a measure of the product mix in the municipality. The DEA method takes into account the case mix when estimating the frontier. However, the distance between the frontier and the average unit behind the frontier might vary with case mix. It is important to remember that since we have no price information on the products (home care and institutionalized care), we do not know which group has the highest total efficiency. Without price information we can only estimate technical efficiency and scale efficiency, not allocative efficiency, which is also a component of total efficiency. Thus, we can make no recommendation of what is better, only point out that the variation of technical efficiency seem to grow with the share of home based nursing.

Share in home care of users in age group…(four age groups) measures how large share of the total population in an age group gets home based nursing services. With the exception of the lowest age group (0-66) all of the parameter estimates are statistically significant and positive. This supports our hypothesis that the higher the coverage of home based nursing, the lower the required resource usage per client. The reasoning is that the nursing sector behaves as if it ranks its potential clients from the ones that require the most nursing to the ones that requires the least, and that it uses this ranking as a prioritized list of which clients to accept first. If the municipality has a larger share of the population in an age group as its clients, we expect the average required resource usage per client to be lower because the average client is healthier.

Share in inst. care of users in age group … (four age groups) is similar to the variables described above, but for institutionalized care. The parameter estimate for the youngest age group (0-66) is positive and statistically significant. It is a priori known that some of these clients require a lot of resource usage, but remember that the number of users in this group (inst. 0-66) is included in the DEA model. It might be that the municipalities who has a relatively large share of these users compared to their total population have healthier clients on average. The only other age group in inst. care that gets a statistically significant parameter estimate is 67-79 where the sign is negative. This is an indication that the "youngest of the oldest" require more resource usage in inst. care than the other groups above 67. It might be that it more for difficult for the clients in this relatively young age group to get inst. care, and that the clients who actually get it require more resources on average than in the older age groups.

Removing the exterior self-evaluators can make a difference. In this case the explained share of the total variance in the model increased as R^2 rose from 17% to 21%. Both coefficient estimates and p-values change, sharpening the estimates of seven coefficients while only three had increased p-values[5]. While numerical changes are small, they are still sizeable considering that only 25 out of 469 observations (5%) were removed. Essentially, we have removed the units that are most likely not to contain any information, i.e. to be pure noise[6].

This is of course not conclusive evidence that one approach is better than the other. The point we want to make is that it *may* make a difference. We have already argued that it makes theoretical sense to remove the exterior self-evaluators. It may be added that in Simar and Wilson (2003) it is conjectured that the bootstrap works better the denser the data. Since we have removed data points in regions that by definition are as "thin" as possible, the bootstrap should also work better. In sum, we feel that we have made a solid case for the advantages of identifying and removing the exterior self-evaluators when doing a two-stage analysis in a DEA setting.

[5] In contrast, excluding all self-evaluators, both interior and exterior, would have lowered R^2 and decreased p-values only for three coefficients and increased them for seven.

[6] Preliminary results from using the homogenous bootstrap suggested by Simar and Wilson (1998) show a standard error of the bias-corrected estimates that is consistently twice as large for the exterior than for the interior self-evaluators, supporting the lack of information content in the efficiency estimates of the former.

5. Conclusions

The units found strongly efficient in DEA studies on efficiency can be divided into self-evaluators and active peers, depending on whether the peers are referencing any inefficient units or not. The contribution of the paper starts with subdividing the self-evaluators into *interior* and *exterior* ones. The exterior self-evaluators are efficient “by default”; there is no firm evidence from observations for the classification. Self-evaluators may most naturally appear at the “edges” of the technology, but it is also possible that self-evaluators appear in the interior. It may be of importance to distinguish between the self-evaluators being exterior or interior. Finding the influence of some variables on the level of efficiency by running regressions of efficiency scores on a set of potential explanatory variables is an approach often followed in actual investigations. Using exterior self-evaluators with efficiency score of 1 in such a “two-stage” procedure may then distort the results, because to assign the value of 1 to these self-evaluators is arbitrary. Interior self-evaluators, on the other hand, may have peers that are fairly similar. They should then not be dropped when applying the two- stage approach.

A method for classifying self-evaluators based on the additive DEA model, either CRS or VRS, is developed. The exterior strongly efficient units are found by running the enveloping procedure “from below”, i.e. reversing the signs of the slack variables in the additive model (1), after removing all the inefficient units from the data set. Which units of the strongly efficient units from the additive model (1) that turn out to be self-evaluators or active peers, will depend on the orientation of the efficiency analysis, i.e. whether input-or output orientation is adopted. The classification into exterior and interior peers is determined by the strongly efficient units turning out to be exterior ones running the “reversed” additive model (9).

The exterior self-evaluators units should be removed from the observations on efficiency scores when performing a two-stage analysis of explaining the distribution of the scores. The application to municipal nursing- and home care services of Norway shows significant effects of removing exterior self-evaluators from the data when doing a two-stage analysis. Thus the conclusions as to explanations of the efficiency score distribution will be qualified taking our new taxonomy into use.

References

Banker, R.D., A. Charnes and W.W. Cooper (1984) "Some models for estimating technical and scale inefficiencies." *Management Science*, 30, pp. 1078-92.

Charnes, A., C.T. Clark, W.W. Cooper, and B. Golany (1985a) "A Developmental Study of Data Envelopment Analysis in Measuring the Efficiency of Maintenance Units in the U.S. Air Forces." *Annals of Operations Research*, 2, 95-112.

Charnes, A., W. W. Cooper, B. Golany, L. Seiford, and J. Stutz (1985b): "Foundations of Data Envelopment Analysis for Pareto-Koopmans Efficient Empirical Production Functions.," *Journal of Econometrics*, 30, 91-107.

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

Cooper, W. W., L. M. Seiford, and K. Tone (2000): *Data Envelopment Analysis. A comprehensive text with models, applications, references and DEA-solver software,* Boston/Dordrecht/London: Kluwer Academic Publishers.

Edvardsen, D. F. and F. R. Førsund (2001): "International benchmarking of electricity distribution utilities, " Memorandum 35/2001, Department of Economics, University of Oslo.

Edvardsen, D. F., F. R. Førsund og E. Aas (2000): " Effektivitet i pleie- og omsorgssektoren" [Efficiency in the nursing- and home care sector], Rapport 2/2000, Oslo: Frischsenteret.

Erlandsen, E. and F. R. Førsund (2002): "Efficiency in the Provision of Municipal Nursing- and Home Care Services: The Norwegian Experience," in K. J. Fox (ed.): *Efficiency in the Public Sector*, Boston/Dordrecht/London: Kluwer Academic Publishers, x-y.

Färe, R. and D. Primont (1995): "Multi output production and duality: Theory and applications," Southern Illinois University at Carbondale.

Farrell, M. J. (1957): "The measurement of productive efficiency," *Journal of the Royal Statistical Society,* Series A, 120 (III), 253-281.

Førsund, F. R. and N. Sarafoglou (2002): "On the origins of data envelopment analysis," *Journal of Productivity Analysis* 17, 23-40.

Nerlove, M. (1965): *Estimation and identification of Cobb – Douglas production functions*, Amsterdam: North-Holland Publishing Company

Seitz, W. D. (1967): "Efficiency measures for steam-electric generating plants", *Western Farm Economic Association, Proceedings 1966,* Pullman, Washington, 143-151.

Seitz, W. D. (1971): "Productive efficiency in the steam-electric generating industry," *Journal of Political Economy* 79, 878-886.

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

Simar, L. and P. W. Wilson (2003): “Estimation and inference in two-stage, semi-parametric models of production processes,” Technical report 0310 IAP statistics network (http://www.stat.ucl.ac.be/Iapdp/tr2003/TR0310.ps).

Torgersen, A.M., F.R. Førsund, and S.A.C. Kittelsen (1996): "Slack-Adjusted Efficiency Measures and Ranking of Efficient Units." *Journal of Productivity Analysis*, 7, 379-39.

Aas, E. (2000): “På leting etter målefeil – en studie av pleie- og omsorgssektoren”, Notater 2000:10, Statistics Norway, Oslo.

CLIMBING THE EFFICIENCY STEPLADDER:

ROBUSTNESS OF EFFICIENCY SCORES IN DEA[*]

by

Dag Fjeld Edvardsen

Norwegian Building Research Institute,

Forskningsveien 3b, NO-0314 Oslo, Norway.

Email: dfe@byggforsk.no

Abstract: The robustness of the efficiency scores in DEA (Data Envelopment Analysis) has been addressed on a number of occasions. It is of crucial importance for the practical use of efficiency scores. The purpose of this paper is to demonstrate the usefulness of a new way of getting an indication of the sensitivity of each of the efficiency scores to measurement error. The main idea is to investigate a DMU's (Decision Making Unit) sensitivity to sequential removal of its most influential peer (with new peer identification as a part of each of the iterations). The Efficiency stepladder approach is shown to provide relevant and useful information when applied on a dataset of Nordic and Dutch electricity distribution utilities. Some of the empirical efficiency estimations are shown to be very sensitive to the validity and existence of one or a low number of other observations in the sample. The main competing method is Peeling, which consists of removing all the frontier units in each step. The new method has some strengths and some weaknesses in comparison. All in all, the Efficiency stepladder measure is simple and crude, but it is shown that it can provide useful information for practitioners about the robustness of the efficiency scores in DEA.

Keywords: DEA, Sensitivity, Robustness, Efficiency stepladder, Peeling.

[*] This study is part of the methodological development within the research project "Productivity in Construction" at the Norwegian Building Research Institute (NBI) financed by the Norwegian Research Council. Finn R. Førsund has followed the whole research process and offered detailed comments. Hans Bjurek, Håkan Eggert, Lennart Hjalmarsson, and Sverre A.C. Kittelsen have also given valuable comments. Any remaining misunderstandings are solely this author's responsibility.

1. Introduction

The robustness of the efficiency scores in DEA has been addressed in a number of research papers. There are several potential problems that can disturb precise efficiency estimation, such as sampling error, specification error, and measurement error. It is almost exclusively the latter that is dealt with in this paper.

It has been proven analytically that the DEA efficiency estimators are asymptotically consistent given that a set of assumptions is satisfied.[1] The most critical assumption might be that there are no measurement errors. The DEA method estimates the production possibility set by enveloping the data as close as possible, in the sense that the frontier consists of convex combinations of actual observations, given that the frontier estimate can never be “below” an observed value. If the assumption of no measurement error is broken we might observe input-output vectors that are outside the true production possibility set, and the DEA frontier estimate will be too optimistic. Calculating the efficiency of a correctly measured observation against this optimistic frontier will lead to efficiency scores that are biased downwards. In other words, even symmetric measurement errors can produce efficiency estimates that are too pessimistic. It is of crucial importance for the practical use of the efficiency scores that information about their sensitivity is available.

The reason why measuring sensitivity is a challenge is in a sense related to the difficulty with looking at n-dimensional space. In two dimensions, and possibly three, one can get an idea of the sensitivity of one observation efficiency score by visually inspecting a scatter diagram. But when the number of dimensions is higher than three, help is needed. The Efficiency stepladder method introduced in this paper is an offer to empirically oriented DEA applications.

This paper is not about detecting outliers; it is about investigating the robustness of each DMUs efficiency score. The main inspiration is Timmer (1971), and the intention is to offer a crude and simple method that works relatively quickly and is available to practitioners as a freely downloadable software package.

In the following only DEA related approaches are considered. There are mainly two ways sensitivity to measurement error in DEA has been examined: (1) perturbations of the

[1] See Banker (1993) and Kneip et al. (1998) for details.

observations, often with strong focus on the underlying LP model, and (2) exclusion of one or more of the observations of the dataset.

The Efficiency stepladder is based on the latter alternative. The main idea is to examine how the efficiency score of a given inefficient DMU develops as the most influential other DMU is removed in each of the iterative steps. The first step is to determine which of the peers whose removal is associated with the largest increase in the efficiency score. This peer is permanently removed, and the DEA model is recalculated giving a new efficiency score and a new set of peers. The removal continues in this fashion until the DMU in question is fully efficient. This series of iterative DMU exclusions provides an "efficiency curve" of the increasing efficiency values connected with each step.

There are few alternative approaches available that provide information about the sensitivity of efficiency scores. Related methods in the literature are Peeling (Barr *et al.*, 1994), Efficiency Order (Sinuany-Stern *et al.*, 1994) and Efficiency Depth (Cherchye *et al.*, 2000). Peeling consists of removing all the frontier units in each step. There are also similarities between the Efficiency stepladder and the Efficiency Order/Efficiency Depth methods. The main difference is that the Efficiency stepladder approach is concerned with the stepwise increase in the efficiency scores after each iterative peer removal, while the Efficiency Order/Efficiency Depth methods are more concerned with the number of observation removals that is required for the DMU in question to reach full efficiency.

The empirical application is mainly used as an illustration on how the Efficiency stepladder method works on real world data. The application is used to show what kind of analysis can be performed using this method. To carry out a full scale empirical analysis is an extensive undertaking, and is outside the scope of this paper.

The layout of the rest of the paper is according to the following plan. Section 2 gives a brief survey of some of the literature related to the sensitivity of the efficiency scores in DEA. Section 3 explains the basic properties of the DEA method. Introduction of the Efficiency stepladder approach is the topic of Section 4. In Section 5, model specification and the basic facts about the dataset are presented. The empirical results and how the Efficiency stepladder method can provide insight about the sensitivity of the dataset used are found in Section 6. Section 7 rounds off the paper with the conclusions.

2. Sensitivity in DEA – a brief survey

The topic of this paper is the sensitivity of the efficiency scores in DEA. Other non-parametric approaches are claimed to be more robust to noisy data. One example is the Order-

M frontier method. It is described in Cazals *et al.* (2002). One application of this method (on U.S. Commercial Banks) is Wheelock and Wilson (2003). Instead of measuring performance relative to the unknown (and difficult-to-estimate) boundary of the production set, performance for a given DMU is measured relative to expected maximum output among banks using no more of each input than the given DMU. The authors claim that this approach permits a fully non-parametric estimation with a much better rate of convergence than DEA, avoiding the usual curse of dimensionality that plagues traditional non-parametric efficiency estimators.

In the following, only DEA related approaches are considered. There are mainly two ways in which sensitivity to measurement error in DEA has been examined: (1) perturbations of the observations, often with strong focus on the underlying LP model, and (2) exclusion of one or more of the dataset observations. Other alternatives have been used when more information about the uncertainty of one or a few of the dimensions is available.[2]

2.1 Investigations based on perturbations of the data in the LP model

Charnes *et al.* (1985) examined the consequences of varying one of the output variables. The intention was to identify the efficient DMUs that have wide ranging effects and distinguish them from others whoose effects are more limited. In the conclusion they state that "*More work needs to be done to extend this for studying the consequences of altering several outputs simultaneously. Input variations and also simultaneous variations of inputs and outputs need to be addressed in other research that should be of value for sensitivity analysis in general*."

One of the papers that picked up that challenge was Charnes *et al.* (1992), who used "distance" (the norm of a vector) in order to determine the "radii of stability" for a DMU. Within this region, data variations do not alter a DMU's status from inefficient to efficient (or vice versa). This is done by centring a box on the original observation for the DMU in question. This box (they refer to it as a "Unit ball", even when it is not round in any possible sense) is defined by the Chebyshev norm which is described by the smallest distance from the centre of the box to any of the sides. For an inefficient DMU the radius defining this box is increased from zero until an observation within this box can be reclassified from inefficient to efficient. The sensitivity of the efficient units is estimated in a similar way.

[2] See Kittelsen *et al.* (2001).

Thompson *et al.* (1994) wanted to determine the magnitudes of data variations that can cause changes in status for the DMUs classified as fully efficient. Their method is based on studying the effects of small increments and decrements in the inputs and outputs with regards to the DMU's classification as efficient or inefficient. They applied this method on two real world datasets (Kansas farming and Illinois coal mining). In the latter they found that within the data variations considered (+/-20% or less in absolute value), 98% of the DMUs in the subset originally classified as 100% efficient were insensitive to potential data errors. The authors claim that their sensitivity analysis shows that DEA results tend to be robust for extreme efficient DMUs.

Zhu (1996) examines how to identify the sensitivity or robustness of efficient DMUs in DEA. His approach is based on linear programming problems whose optimal values yield particular regions of stability. Sufficient and necessary conditions for variations in inputs and outputs of an efficient DMU to maintain full efficiency are provided.

2.2 Investigations based on exclusion of observations from the dataset

An early and influential contribution was Timmer (1971). This paper heavily quoted Farrell (1957), but used deterministic frontiers (estimated with linear programming) instead of DEA. Though not mentioned in its abstract, the paper was a pioneering contribution when it comes to measuring the sensitivity of the efficiency scores when removing selected units from the dataset. Timmer showed two ways to do this. The first alternative he suggested was to remove observations from the dataset until a given percentage of the dataset is outside the probabilistic frontier. The other alternative he suggested was to remove efficient observations one by one until the resulting frontier stabilizes. Timmer claimed that either of these approaches may overcome the objections to estimating a frontier function because of data problems.

Superefficiency was introduced in Andersen and Petersen (1993). It was introduced to rank efficient units, but as pointed out in Banker and Chang (2000), it is probably more useful for detecting outliers when there is reason to believe that the data might be noisy. Superefficiency is a measure of the relative radial distance from the origin to the DMU in question,when the frontier is estimated without this DMU included in the dataset. Superefficiency is by construction greater than (or equal to) one. A superefficiency value of 1.2 implies that the DMU is positioned "20% outside" where the frontier would have been without this DMU (in a radial sense).

Peeling is described in Barr *et al.* (1994). This approach measures how much the efficiency of a DMU would change if the whole frontier was removed. They used the allegory that peeling in DEA is like removing layers from an onion. The DEA dataset can be seen as a series of frontiers inside other frontiers. If we remove all the observations in the first frontier, a new frontier is generated when the LP model is recalculated for the remaining units. This continues until there are no more observations left. With peeling one is typically mostly concerned with which frontier a DMU belongs to -- the one where it becomes efficient. A weakness is that a frontier in DEA typically consists of different numbers of units. For the individual DMU, removing one single unit can be sufficient for it to reach the frontier. Removing the entire frontier is measured as one operation, independently of the number of units this particular frontier consisted of. One attractive aspect with Peeling is that it is very fast to compute. Peeling is well known in the DEA research community, but surprisingly few empirical DEA applications take advantage of this method. One possible explanation is that none of the mainstream commercial and freeware DEA software packages offer automatic generation of the layer number of each DMU. Peeling in DEA is in spirit very close to Timmer (1971), but since Timmer did not use DEA the selection of which and how many DMUs to remove is a little different. Timmer suggested removing a given number or a given percentage of the DMUs, while Barr et al. suggest removing the entire frontier – independently of whether the frontier is made up of 1 or 20 DMUs.

Sinuany-Stern *et al.* (1994) introduced *Efficiency Order* as "*the number of units we need to delete in order to reach efficiency*." What algorithm one should use to identify the number of units that is required to be deleted is not explained in detail. A similar approach can be found in Cherchye et al. (2000),[3] who used a mixed integer algorithm to identify the Efficiency Order. Further information on how the Efficiency Order relates to the Efficiency stepladder is given in Section 4.

Wilson (1995) investigated the consequences of removing observations from the dataset. If removing an observation makes big differences in the efficiency scores of the other DMUs, then the area of the dataset where this input/output combination was found is not densely populated, and convex combinations of other DMU offer little help. This is an indication that the observation in question is a possible outlier and should be investigated for measurement error. By definition this approach works only on the fully efficient units.

[3] They use the term "efficiency depth", and do not refer to Sinuany-Stern *et al.* (1994).

3. Data Envelopment Analysis

3.1 The origins of DEA

The original idea behind DEA was introduced in Farrell (1957). It was further developed in a very influential paper by Charnes, Cooper and Rhodes (1978). The term Data Envelopment Analysis (DEA) was coined in their paper. However, the first use of Linear Programming (LP) in the calculation of the DEA efficiency scores was made by Farrell and Fieldhouse (1962). The DEA model with variable returns to scale is often referred to as the BCC-model (Banker, Charnes and Cooper, 1984), but it was introduced in Afriat (1972) in the single output case, and empirically implemented in the case of multiple outputs in Färe, Grosskopf and Logan (1983).[4]

Banker (1993) proved that the output oriented efficiency score is consistent in the case of a single output, while Kneip *et al.* (1998) showed statistical consistency and rate of convergence in the general multiple-input and multiple-output case. Unfortunately the rate of convergence is low, leading to sampling bias. The expected size of this bias increases exponentially in the number of inputs and outputs for a given sample size. The bias can be estimated and the efficiency estimate bias adjusted with a statistical technique referred to as bootstrapping (Efron, 1979). If the required number of inputs and outputs is large compared to the number of DMUs available, the standard errors for the (bootstrapped) bias corrected efficiency scores will be very large, and the discrimination of the efficiency scores will have little statistical significance. Including too few inputs and outputs will reduce the curse of dimensionality, but will lead to a wrongly specified efficiency model. In a sense this is worse, because the confidence intervals will misleadingly tend to be smaller the fewer inputs and outputs we include.[5] Including too many inputs or outputs (as long as the correct ones are included) will not make the efficiency estimator inconsistent (in an asymptotic sense), but with finite samples it will make the efficiency estimate more noisy and biased. Statistical tools for choosing model specification have been developed, but they do (of course) require that observations of the important inputs and outputs are available, and that a sufficiently large number of DMUs are available for the tests to give significant results (depending of the power of the tests) . This line of thought leads back to Banker (1993, 1996) and Kittelsen (1993).

[4] For the history of the development of DEA, see Førsund and Sarafoglou (2002).

[5] This is a complicated mechanism, and will not be covered in further detail in this paper.

3.2 The LP formulation of the DEA model

Førsund and Hjalmarsson (1979) define the measures E_1 to E_5, where E_1 is radial efficiency assuming variable returns to scale. This is the same as the Banker *et al.* (1984) model formulated as:

$$
\begin{aligned}
& E_{1i} \equiv Min\,\theta_i \\
& s.t. \\
& \sum_{j \in N} \lambda_{ij} y_{mj} - y_{mi} \geq 0, m = 1,\dots, M \\
& \theta_i x_{si} - \sum_{j \in N} \lambda_{ij} x_{sj} \geq 0, s = 1,\dots, S \\
& \sum_{j \in N} \lambda_{ij} = 1 \\
& \lambda_{ij} \geq 0, j \in N
\end{aligned}
\tag{1}
$$

The usage of symbols in model (1) is as follows: E_{1i} is the input saving VRS efficiency for DMU I, θ_i is a scalar, S is the number of inputs dimensions, M is the number of output dimensions, and N is the set of DMU. The indices i and j belong to the set N, y_{mj} is the level of output and x is the level of intput. λ_{ij} is a reference weights.

The peers for DMU_i in problem (1) are the ones for which λ_{ij} is strictly positive. If DMU_i is strongly efficient then λ_{ii} has the value of 1. In other words, a strongly efficient DMU is its own peer.[6] Notice that not all units with radial efficiency equal to 1 are Pareto efficient. They might have slack in one or more dimensions. To identify which of the DMUs are Pareto efficient, the “additive” DEA model can be used. Here the sum of slacks for each DMU is maximized, and only the Pareto efficient units have zero slack (see Charnes *et al.*, 1985).

Figure 1 is an illustration of how the DEA model works in the VRS case with two inputs and one output. The DMUs A, B and C are efficient and define the boundary of the Production Possibility Set (PPS), while DMU D is inefficient and is positioned strictly in the interior of the PPS. The input saving radial efficiency of DMU D is equivalent to the proportional radial contraction of all inputs possible while staying within the PPS. The radial contraction is stopped at point F. The radial input saving efficiency of DMU D (E_1) is equal to the ratio GF/GD. However, point F is not a Pareto efficient point. In addition to reducing both

[6] If two (or more) DMUs have identical input-output vectors, the choice of peer(s) is not unique.

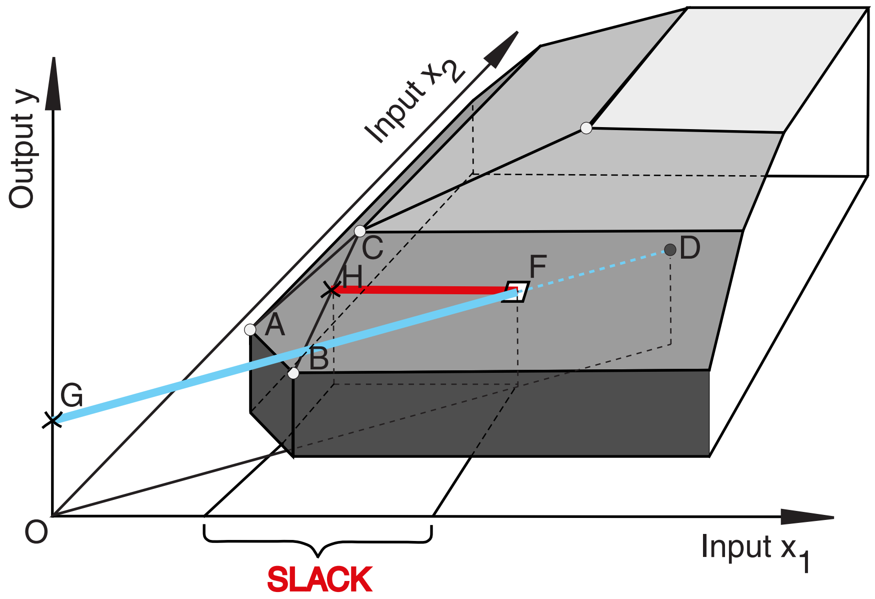

Figure 1: Illustration of DEA with two inputs and one output (VRS) using the radial Farrell input reducing efficiency measure.

the inputs with the same percentage, it would also be possible to further reduce the usage of input x_1 equivalent with the distance FH. The existence of this extra slack is not captured by the E_1 measure.

4. The Efficiency stepladder – a method for measuring sensitivity in DEA

The basic idea behind the Efficiency stepladder approach is quite similar to the efficiency order (Section 2.2). The robustness of the efficiency score of the unit under investigation is examined in light of the exclusion of other observations in the sample. In both cases one is interested in the lowest possible number of observations that has to be removed for the DMU in question to reach the frontier, but with the Efficiency stepladder approach there is greater focus on the entire development from the original efficiency and then step by step until the DMU is fully efficient. The exact algorithm used is presented in Section 4.1, but the basic idea used in the computer program is in each step to determine which of the peers whose removal is associated with the largest efficiency increase. This peer is then removed, and the DEA model is recalculated leading to a new peer group set. This is repeated until the DMU has an efficiency score of 100%.

One alternative algorithm could be to iterate over all alternatives and then determine which sequence of observation removals that most quickly moved the DMU to the frontier. This approach would however extremely time consuming with even medium sized datasets because of the very high number of possible sequences to consider. A natural implementation when an unlimited number of CPU cycles are not available is a first step optimal algorithm. It can easily be shown that this algorithm can choose stupid paths when digging out peers in order to get the unit to the frontier with as few steps as possible (one example is provided further down in the text), but the results are still useful as long as one remembers that a high number of steps to the frontier should not be taken as very strong evidence that the DMU is inefficient. On the other hand a low number of Efficiency stepladder iterations before we reach the frontier means with certainty that the efficiency of the DMU in question is very sensitive to the quality of observation for the DMUs removed in the Efficiency stepladder sequence. In other words, be concerned if the slope of the Efficiency stepladder is steep, but don’t be too calm if the increase is slow.

The computer program used to calculate the numbers in this paper is “DagEA” which has been developed for this exact purpose. In the current version it is a front end[7] that uses DEAP (Coelli, 1996) as its DEA solver. Both DagEA will be and DEAP are freely available on the Internet. In the front end of DagEA there are routines for automatic calculation of the Efficiency stepladder for a dataset.[8] For middle-sized datasets, calculation is relatively fast, but calculation time increases exponentially with the dimensionality.[9]

4.1 The Efficiency stepladder approach illustrated

The one step optimal algorithm is very simple:

1. Calculate the DEA efficiency score for the unit of interest (“DMU P_1”) and write down which units serve as peers for DMU. The peers are characterized by having a strictly positive λ in the optimal solution of the LP model formulated in (1).

[7] A front end is a computer program that provides the visual interface that the user interacts with, but it uses another computer program as its calculation engine.

[8] DagEA is developed by Dag Fjeld Edvardsen. It will eventually be downloadable for free from http://home.broadpark.no/~dfedvard.

[9] Calculating all the Efficiency stepladder values on the dataset used in this paper took 37 minutes on a 1.2 Ghz Pentium M notebook, but there are still possibilities for optimizing the source code. Increasing the speed by a factor of 10 on the same hardware might be realistic.

2. For each of the peers identified in the step above: Calculate the efficiency score of the DMU P_1 if that peer is removed from the dataset, write down the efficiency score, and put the peer back in the dataset before the efficiency of DMU P_1 is calculated with another peer temporarily removed.
3. Permanently remove the most influential peer identified in the step above. This is the one that has experienced the largest change in efficiency associated with its removal (this is equivalent with the peer whose removal makes the efficiency score of DMU P_1 the largest). This efficiency score and the most influential peer's identity are added to the Efficiency stepladder table.
4. Repeat (1)-(3) while permanently removing the peers identified in (3) and for each iteration adding the id-number of the most influential peer and the efficiency score associated with its removal. Stop the repetitions when the efficiency of DMU P_1 reaches 1.

4.2 Possible problems with the one step optimal algorithm

Figure 2 is an example of how a one step optimal routine can potentially choose a route towards the frontier that takes a higher number of steps than necessary. The challenge is to find the shortest number of sequential peer exclusions that will make DMU J fully efficient. Looking at Figure 2, it is easy to see that excluding DMU H and then DMU I results in DMU J reaching the frontier in two steps. However, the one step optimal algorithm by definition

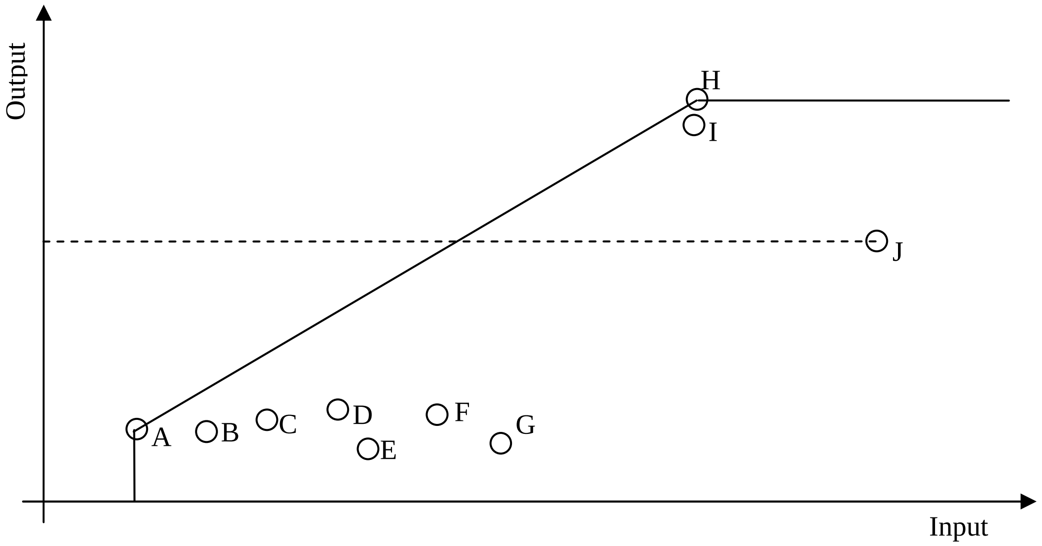

Figure 2: Illustration of how the one step optimal algorithm can choose the wrong path.

only compares the alternative one step further down the road. Because of this it will choose to remove DMU A instead of removing DMU H since this will result in the greatest increase in the efficiency score for DMU J. Next, with DMU A out of the way, it will choose to remove DMU B instead of DMU H, for the same reason. This continues as the one step optimal algorithm chooses to remove the DMUs C, D, E, F, and G. Only after this it decides to eliminate DMU H and DMU I. In this example the algorithm used nine steps to accomplish what really needed only two steps.

However, this example is constructed to show the one step optimal algorithm in the worst possible light. There is no indication that such behaviour is common when real world data is used. At the same time it is a demonstration of why it is important to think of the Efficiency stepladder approach as one way safe. If it reports that it only takes a low number of sequential peer removals to move from large inefficiency to the frontier one can be certain that the sensitivity of the efficiency score is high. But as demonstrated by Figure 2, one should not be too calm if the algorithm indicates that a high number of peer removals are necessary. It could be tempting to use brute force and compare all possible peer removals until the path with the lowest number is found, but this may not be a practical alternative because this exhaust the capacity of today's generation of PCs. The reason is that the number of alternatives to compare will easily be extremely high.

There are some ways to reduce this problem. One way is to cluster two observations close to each other into a singly entity, and possibly minimize a penalty function where removing this entity counts as double the removal of only one DMU. Another possibility is to combine the Efficiency stepladder approach with peeling, and notice the cases where there are large differences between the results of these two methods. Peeling has it own weaknesses, but it is simpler, faster and in some cases more robust in difficult situations such as the one presented in Figure 2.

4.3 Efficiency stepladder for fully efficient units

The Efficiency stepladder can also be calculated for fully efficient DMUs, or DMUs which become fully efficient after having gone through a number of iterations from their original position below the frontier. In these cases we measure using the "superefficiency" concept (Andersen and Petersen, 1993). We continue to do Efficiency stepladder iterations, and stop when the superefficiency is undefined. The higher the number of steps from the original position to undefined, the more units involved in calculating the efficiency of the DMU. The reason why the sensitivity of the fully efficient DMUs is interesting is the same as for the inefficient units – it is relevant to know how robust the frontier that this unit is compared with is to measurement error. If the efficiency of the unit becomes undefined after a low number of Efficiency stepladder iterations, this is an indication that the part of the frontier that this unit is compared with is not very robust. Calculating the Efficiency stepladder involves removing the most influential peer for a DMU in each step. The fully efficient DMUs are their own peers, and when we remove them the value is by definition greater than or equal to 1 (they are no longer allowed to be part of the convex combinations that define the frontier, but remain as ghost units that we measure against). For this reason ESL(1) for the frontier units is the same as superefficiency, while ESL(2) and later take the superefficiency concept further.

5. Model specification and data

The empirical part of this paper is mainly intended as an illustration of the ESL method. The dataset is a quite typical example of the datasets used in empirical applications for the DEA method when it comes to the number of observations and the number of inputs and outputs. The dataset is cross section data on the Nordic and Dutch electricity distributors in 1997 (see Edvardsen and Førsund, 2003). The data was collected by the national regulators.

The key characteristics of the data are presented in Table 1. The difference in size between the DMUs is large, as revealed by the last two columns. *TOM* is Total Operating and Maintenance cost (including labor costs) measured in Swedish kronor in thousands. *LossMWH* is energy loss in megawatt hours, *RV* is Replacement Value measured in Swedish kronor in thousands. *NumCust* is the number of costumers. *TotLines* is the total length of lines. *MwhDelivered* is the sum of megawatt hours delivered. See Edvardsen and Førsund (2003) for further details on the content and history of the dataset.

Table 1. Summary statistics. Cross-section 1997. Number of units 122.

	Average	Median	Standard Deviation	Minimum	Maximum
TOM(kSEK)	152388	97026	182923	11274	981538
LossMWh	91449	52318	104777	7020	615281
RV (kSEK)	2826609	1907286	3288382	211789	22035846
NumCust	109260	55980	163422	20035	1052096
TotLines	7640	4948	8824	450	54166
MWhDelivered	2110064	1003472	2815025	166015	178054730

6. The results

6.1 The basic DEA efficiency results

Figure 3 is an *Efficiency Diagram*[10] showing the results of the efficiency calculations assuming Variable Returns to Scale (VRS). Each of the efficiency scores is calculated by solving the linear programming problem in (1). The DEA calculations shown in Figure 3 assume no measurement error, but what if this assumption does not hold? The purpose of the Efficiency stepladder approach is to examine the sensitivity of efficiency scores to measurement errors.

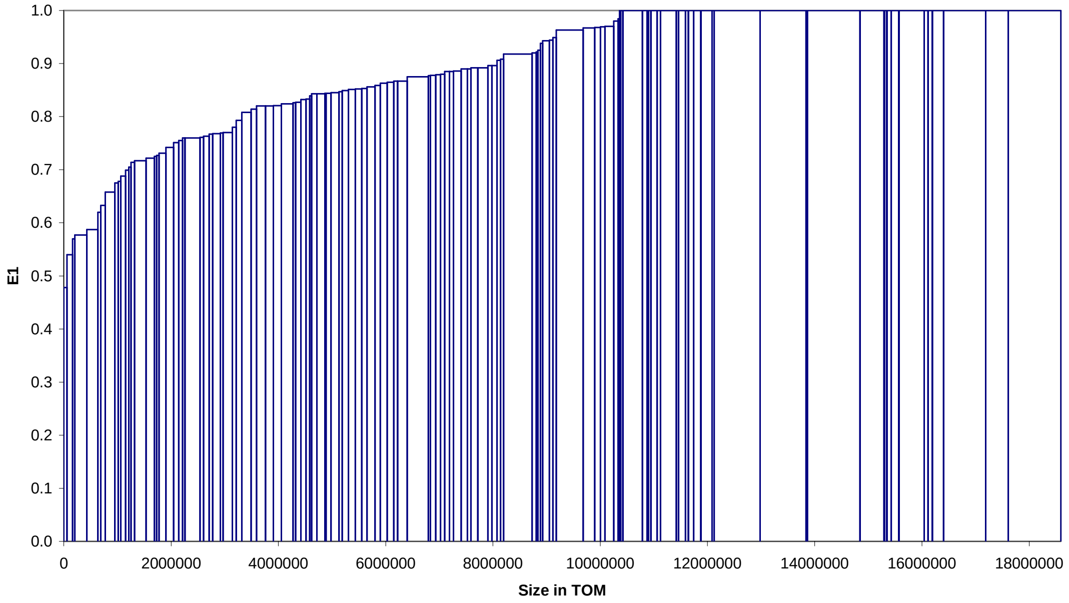

Figure 3: Input saving efficiency scores when assuming VRS (E_1).

[10] One interesting feature of Efficiency diagrams is that both the heights and the widths of the bars can contain information – unlike a bar chart where only the heights of the bars are actively used. This is especially useful when illustrating the results of efficiency analysis. The efficiency of each DMU is shown by the height of the bar, while its economic size (man-years in Fig. 3) is shown by the width of the bars. This means that it is possible to examine whether there are any systematic correlations between the sizes of the units and their efficiencies. Another interesting geometric aspect of these figures is that they are sorted according to increasing efficiency from left to right. The distance from the top of each bar to 1.00 is a measure of that DMU's inefficiency, and the width of the bar is a measure of its economic size. For this reason the area above each of the bars is proportional to the economic cost of that DMU not being 100% efficient. This means that there will typically be a "white triangle" above the inefficient units, and that the size of that area is proportional to the economic cost of the total inefficiency in the sample.

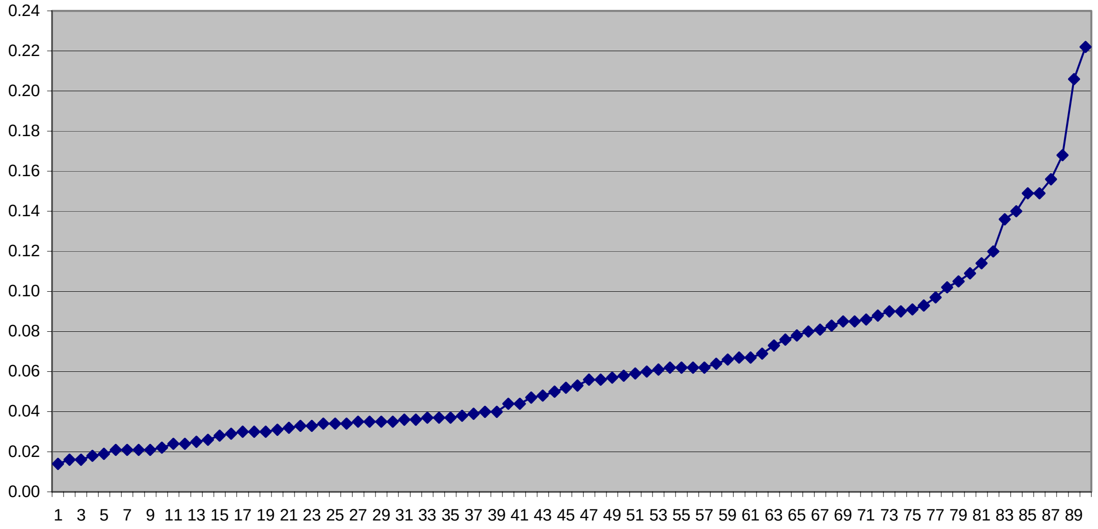

Figure 4: First step ESL values for all the inefficient DMUs, sorted in increasing order.

The changes in the efficiency scores from the first step in the ESL algorithm are shown in Figure 4. For more than half of the inefficient DMUs the changes after removing their originally most influential peer, ESL(1), is larger than 5 percentage points, and for a fifth of the DMUs the changes are larger than 10 percentage points. Two DMUs experience increases in their efficiency scores larger than 20 percentage points. This suggests that the individual efficiency scores in DEA applications strongly depend on the assumption of no measurement error. If the most influential of its peers is outside the true production possibility set, one can get a very large negative measurement-error bias in the estimated efficiency scores.

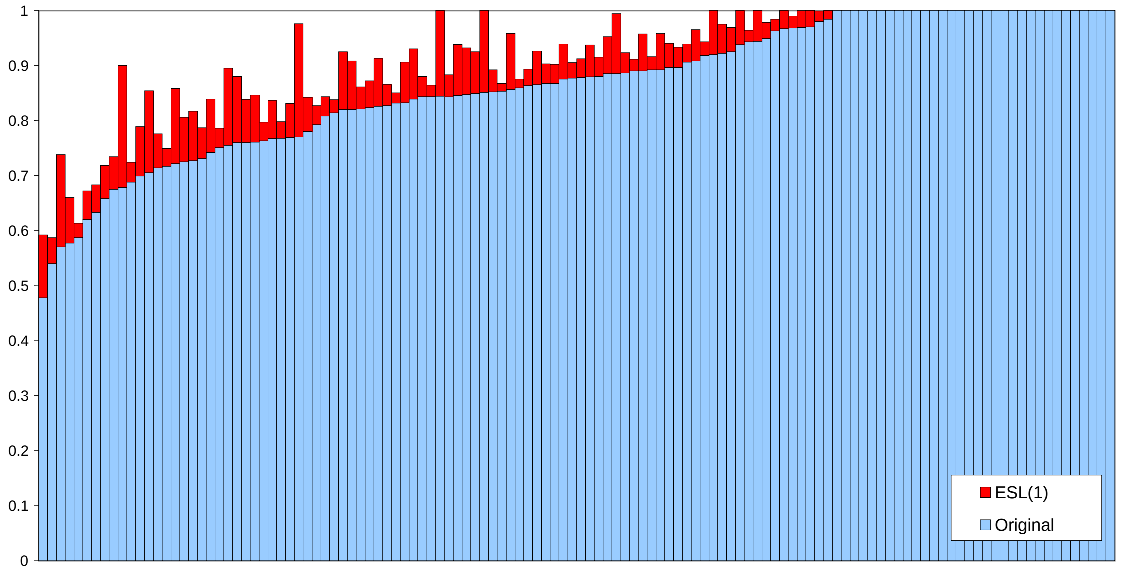

Figure 5: The original efficiency scores and the ESL(1) values for all the inefficient DMUs, sorted pairwise.

Figure 5 is similar to Figure 4, but now the ESL(1) values for the inefficient units are shown together with their DMU’s original efficiency (sorted pairwise). A visual inspection of Figure 1 confirms that a number of the inefficient units move from being quite inefficient to being quite efficient (if we place the border between these two conditions at the *ad hoc* value of 0.85). It is also interesting to notice that there does not seem to be any strong pattern concerning the correlation between the original value and the ESL(1) value, especially if one sees it in light of the DMUs that are originally assigned a high efficiency being limited in how big the ESL(1) value can be since the efficiency number can not be larger than 1.

Figure 6 is similar to Figure 5, but shows the Efficiency stepladder values for the first six steps of the sequential Efficiency stepladder iterations for all the inefficient DMUs. The changes in the efficiency score with ESL(2) tend to be smaller than in ESL(1), but there are examples of the opposite. A few of the DMUs have low sensitivities to the validity of their peers, but the general picture is that most of the DMUs experience large changes in their efficiency scores after two or three Efficiency stepladder iterations.

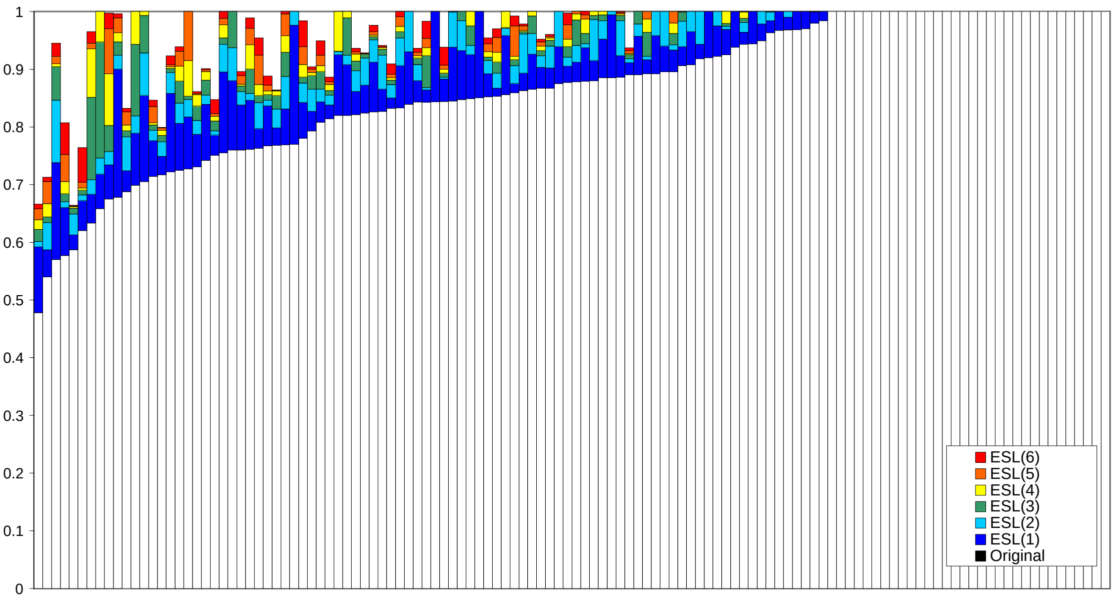

Figure 6: Stacked bar chart showing original efficiency and ESL(1) to ESL(6).

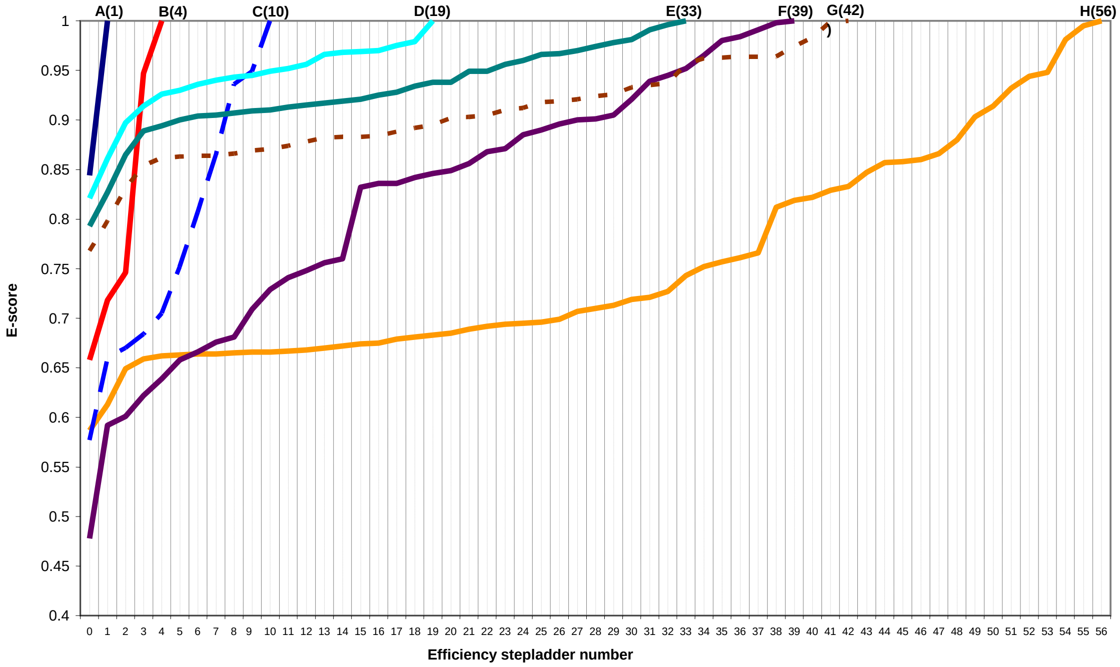

Figure 7: The Efficiency stepladder for a few selected inefficient DMUs (horizontal axis is the Efficiency stepladder number).

Figure 7 shows the ESL curves for seven of the inefficient DMUs in the dataset (referred to as A-H). They are selected because their curves show some of the different developments. By construction all of the curves are non-decreasing. The identity of each of the curves is indicated on the top of the figure, together with the required number of Efficiency Stair iterations for that DMU to reach full efficiency. DMU A has an original efficiency of 0.85, but it becomes fully efficient after excluding only one of its peers from the sample. DMU B has an original DEA efficiency score of 0.66, but after four steps it reaches the frontier. DMU C starts at 0.57, but has a very steep increase in efficiency. The other DMUs (E-G) have slower efficiency increases. One natural summary measure of the steepness of the Efficiency stepladder is the average increase per step. This is presented in Table 2. It is worth noticing that DMU B starts at a much lower efficiency than DMU D, but because of DMU B’s lower average increase per step it reaches the frontier in only four steps while DMU D needs 19 steps. DMU H is a bit different from the other DMUs in that it experiences a mostly convex (broadly speaking) development, while the other DMUs that experience a large number of ESL steps before they reach the frontier (D-G) follow a mostly concave pattern.

Table 2: Average increase per step for the selected DMUs

	Original value	Inefficiency	Steps	Average increase per step
A	0.84	0.16	1	0.16
B	0.66	0.34	4	0.085
C	0.57	0.43	10	0.043
D	0.78	0.22	19	0.012
E	0.79	0.21	33	0.006
F	0.47	0.53	39	0.013
G	0.77	0.23	42	0.005
H	058	0.42	56	0.008

6.2 Efficiency stepladder (ESL) versus Efficiency Order

The Efficiency Order approach (Sinuany-Stern *et al.*, 1994) is mainly concerned with the minimum number of DMUs that need to be deleted for the DMU in question to reach the frontier. In relation to Figure 8 the efficiency order approach would be interested in the number of steps each of the DMUs required to get to the frontier, while the ESL approach is more focused on the steepness of the Efficiency stepladder. Another difference is that the Efficiency Order approach does not seem to be interested in the first few steps in the stepladder, but only in the total number of iterative DMU exclusions that leads to full efficiency. This might be a weakness of the Efficiency Order approach since the likelihood of a low number of observations to be outside the true production possibility may be low. Correspondingly, the likelihood of a large number of observations to be infected with serious measurement errors may be quite low. In other words, one of the most interesting indications of a largely inefficient DMU's robustness is its sensitivity to one, two, or maybe three sequential peer removals. The Efficiency Order approach would not capture the fact that the first peer removal might move an inefficient unit from 40% to 90% efficiency if the rest of its way to full efficiency takes 10 more steps.

The ESL approach proposed in this paper is concerned with the changes in efficiency in each step, and not only with the number of removals necessary before the frontier is reached. Another difference is that the ESL approach is also relevant for the fully efficient units. The first step is then identical to the Superefficiency (Andersen and Petersen, 1993).

Thirdly, to calculate the Efficiency Order using the algorithm proposed in Cherchye *et al.* (2000), specialised software and some knowledge of computer programming are required. The authors claim that the calculation of the Efficiency Order (they refer to it as "efficiency depth") with their approach "*should not involve substantial computational burden*" and "*require only minimal effort using an ordinary PC desktop* (sic)." The exact approach for calculating the efficiency depth is unclear, and they formulate a Mixed Integer Linear Programming (MILP) problem without explaining how much CPU time that is required on a desktop PC. Identifying which of the DMUs should be removed and in what sequence is left to the CPLEX[11] MILP optimizer. It is not certain that a different MILP optimizer would choose the same path towards the frontier.

The simpler algorithm used in this paper is more accessible to practitioners. A computer program that calculates the Efficiency stepladder has been developed to calculate the numbers presented. It will be freely available on the Internet so that practitioners can use it to get some crude but useful information on the sensitivity of the efficiency scores in DEA. Since the algorithm always chooses the one-step optimal solution it is predictable how the Efficiency stepladder is constructed.[12]

7. Conclusions

Ideally sensitivity analysis, detection of potential outliers, and estimation of sampling bias should be carried out simultaneously. It is easier to detect outliers if we have some information about the sampling bias, and it is easier to estimate sampling bias if we have first identified the outliers. There have been developments made on all these areas in the last few years, but at the time of writing no single method offers a solution to all the mentioned challenges.

The Efficiency stepladder method is simple and crude, but it can still be useful for applied DEA investigations. It should be thought of as one way safe: An Efficiency stepladder that is very steep is a clear indication that the DEA estimated efficiency is strongly dependent on the correctness of a low number of other observations. A slow increase on the other should

[11] CPLEX is a commercial optimizer capable of solving Mixed Integer Linear Programming problems. See http://www.ilog.com/products/cplex/ for more information.

[12] But even with the open algorithm used in the ES approach there can be situations where the computer program has to choose between two or more equally good alternatives. However, in most of these cases this is the last step before the DMU in question reaches the frontier, and the "problem" is that removing any of several alternative peers will lead to full efficiency. In this case the Efficiency stepladder curve and all the efficiency values remain the same; only the identity of the last peer removed will differ.

not be interpreted as a strong indication that the efficiency is at least this low. The reason is that the method is only one-step-optimal. In addition to measuring the sensitivity of the e-scores for efficient and inefficient units, it might be used in combination with bootstrapping to identify possible outliers. The necessary software for carrying out the Efficiency stepladder calculations will be made available from the author's website.

The purpose of the ESL method is to examine the sensitivity of the efficiency scores for measurement errors. Bootstrapping on the other hand is in the DEA context (primarily) used to measure sensitivity to sampling errors. We would expect that a DMU with a large ESL(1) value would also have a large standard error of the bias corrected efficiency score. The reason is that we expect the part of the (input, output) space where the DMU is located to be sparsely populated.

Tentative runs have shown statistically significant and positive correlation between the ESL(1) values and the standard errors of the bootstrapped bias corrected efficiency scores. Furthermore, there is strong empirical association between the ESL(1) values for the fully efficient DMUs (=superefficiency) and the sampling bias estimated using bootstrapping. This is a promising topic for further research.

References

Afriat, S., 1972, Efficiency estimation of production functions, *International Economic Review*, 13(3), 568-598.

Andersen, P. and Petersen, N.C., 1993, A Procedure for Ranking Efficient Units in Data Envelopment Analysis. *Management Science*, 39(10), 1261-1264.

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.

Banker, R.D., A. Charnes and W.W. Cooper, 1984, Some Models for Estimating Technical and Scale Inefficiencies in Data Envelopment Analysis, *Management Science* 30, 1078-1092.

Banker, R.D. and Chang, H., 2000, Evaluating the Super-Efficiency Procedure in Data Envelopment Analysis for Outlier Identification and for Ranking Efficient Units, Working Paper from the University of Texas at Dallas.

Barr, R.S., M.L. Durchholz and Seiford, L., 1994, Peeling the DEA Onion. Layering and Rank-Ordering DMUs Using Tiered DEA, Southern Methodist University technical report, Dallas, Texas.

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

Charnes, A., Haag, S., Jaska, P., and Semple, J.,1992, Sensitivity of Efficiency Calculations in the Additive Model of Data Envelopement Analysis, , *International Journal of System Sciences*, 23, pp. 789-798.

Charnes, A., Cooper, W.W. and Rhodes, E., 1978, Measuring the efficiency of decision making units, *European Journal of Operations Research* 2, 429-444.

Charnes, A., Cooper, W.W., Lewin, A.Y. , Morey, R.C., and Rousseau, J.J.., 1985. Sensitivity and Stability Analysis in DEA. *Annals of Operations Research* 2 139-150.

Cherchye, L. Kuosmanen, T. and Post, G.T., 2000, New Tools for Dealing with Errors-In-Variables in DEA, Katholike Universiteit Leuven, Center for Economic Studies, Discussion Paper Series DPS 00.06.

Coelli, T.J. 1996, “A Guide to DEAP Version 2.1: A Data Envelopment Analysis (Computer) Program”, CEPA Working Paper 96/8, Department of Econometrics, University of New England, Armidale NSW Australia.

Edvardsen, D.F. and Førsund, F.R. 2003: International benchmarking of electricity distribution utilities, *Resource and Energy Economics*, 25, 353-371.

Farrell, M.J.,1957, The measurement of productive efficiency, *J.R. Statis. Soc*. Series A 120, 253-281.

Farrell, M.J. and Fieldhouse M., 1962, Estimating efficient production functions under increasing returns to scale, J.*R. Statis. Soc*. Series A 125, 252-267.

Førsund, F.R. and N. Sarafoglou, 2002, On the origins of Data Envelopment Analysis, *Journal of Productivity Analysis* 17, 23-40.

Färe, R., Grosskopf, S. and Logan, J., 1983, The relative efficiency of Illinois electric utilities, *Resource and Energy* 5, 349-367.

Kittelsen, S.A.C., 1993, Stepwise DEA; Choosing Variables for Measuring Technical Efficiency in Norwegian Electricity Distribution, Memo 06/1993 Department of Economics, University of Oslo

Kittelsen, S.A.C., G.G. Kjæserud and O.J. Kvamme: Errors in Survey Based Quality Evaluation Variables in Efficiency Models of Primary Care Physicians, HERO Memoranda 24/2001, Oslo. [http://www.oekonomi.uio.no/memo/memopdf/memo2401.pdf]

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

Sinuany-Stern, Z., A. Mehrez and A. Barboy, 1994, Academic Departments Efficiency via DEA, *Computers Ops. Res*., vol. 21, No. 5, pp. 543-556.

Timmer, C.P., 1971, Using a Probibalistic Frontier Production Function to Measure Technical Efficiency, *Journal of Political Economy*, Vol. 79, No. 4 (Jul. – Aug. 1971), 776-794.

Thompson R., Dharmapala, P.S. and Thrall, R.M., 1994, Sensitivity analysis of efficiency measures with applications to Kansas farming and Illinois coal mining. In: Data Envelopment Analysis, Theory, methodology and applications. Edited by Charnes A., W. Cooper, Lewin A.Y, Seiford, L.M. Kluwer.

Wilson, P.W., 1995,Detecting Influential Observations in Data Envelopment Analysis, *Journal of Productivity Analysis* 6, 27-46.

Zhu, J., 1996, , Robustness of the efficient DMUs in data envelopment analysis, European *Journal of Operations Research*, 90, pp. 451-460.

EFFICIENCY OF NORWEGIAN CONSTRUCTION FIRMS[*]

by

Dag Fjeld Edvardsen

Norwegian Building Research Institute,
Forskningsveien 3b, NO-0314 Oslo, Norway.
Email: dfe@byggforsk.no

Abstract: Efficiency studies of the construction industry at the micro level are few and far between. In this paper information on multiple outputs is utilized by applying Data Envelopment Analysis (DEA) on a cross section dataset of Norwegian construction firms. Bootstrapping is applied to select the scale specification of the model. Constant returns to scale was rejected. Furthermore, bootstrapping was used to estimate and correct for the sampling bias in the DEA efficiency scores. One important lesson that can be learned from this application is the danger of taking the efficiency scores from uncorrected DEA calculations at face value. A new contribution is to use the inverse of the standard errors (from the bias correction of the efficiency scores) as weights in a regression to explain the efficiency scores. Several of the hypotheses investigated are found to have statistically significant empirical relevance.

Keywords: Construction industry, DEA, efficiency, bootstrapping, weighted two stage.

[*] This study is part of the research project “Productivity in Construction” at the Norwegian Building Research Institute (NBI) financed by the Norwegian Research Council. I would like to thank Thorbjørn Ingvaldsen (NBI) for insightful comments about the nature of construction activities. Finn R. Førsund (University of Oslo) has followed the whole research process and offered detailed comments. Sverre A.C. Kittelsen (Frisch Centre) has been invaluable in the development of the software for bootstrapping, and for giving detailed comments. Comments from Lennart Hjalmarsson, Håkan Eggert and Hans Bjurek have highly improved the presentation of this paper. Roger Jensen (Statistics Norway at Kongsvinger) provided the dataset and offered help in the data selection process. Any remaining errors are solely this author’s responsibility.

1. Introduction

Low productivity growth of the construction industry in the nineties (based on national accounting figures) is causing substantial concern in Norway. To identify the underlying causes investigations at the micro level are needed. However, efficiency studies at the micro level of the of the construction industry are very rare.[1]

The objective of this study is to analyze productive efficiency in the Norwegian construction industry. A piecewise linear frontier is used, and technical efficiency measures (Farrell, 1957) are calculated on cross section data following a DEA (data envelopment analysis) approach (Charnes *et al.*, 1978).

The DEA efficiency scores are bias corrected by bootstrapping (Simar and Wilson, 1998, 2000), and a bootstrapped scale specification test is performed (Simar and Wilson, 2002). A new contribution is to use weights based on the standard errors from the bootstrapped bias correction in the two stage model when searching for explanations for the efficiency scores.

One reason for the small number of efficiency analyses of the construction industry may be the problem to "identify" the activities in terms of technology, inputs and outputs in this industry. It is well known that there are large organizational and technological differences between building firms. Even when the products are seemingly similar there are large differences in the way projects are carried out. For instance some building projects use a large share of prefabricated elements, while other projects produce almost everything on the building site. This often happens even when the resulting construction is seemingly similar. It is interesting to note that projects with such large differences in the technological approach can exist at the same time. Moreover, the composition of output varies a lot between different construction companies so the definition of the output vector may also be a problem. Thus to capture such industry characteristics, a multiple input multiple output approach is required.

A debated issue is whether an efficiency analysis should be carried out at the project level or at the firm level. In many ways it is more natural to think of the project level as the decision making unit (DMU) in this industry. In addition it might be easier to find relatively homogenous projects than firms. A third aspect is that when one tries to explain any

[1] Two Scandinavian studies are Jonsson (1996) which looked at construction productivity at the project level and Albriktsen and Førsund (1991) which investigated the efficiency of Norwegian construction firms. The latter was based on a parametric frontier approach specifying only one output.

efficiency differences it is likely that there are bigger differences between the projects than between the firms when it comes to choice of construction technology. However, the required data for studying productivity at the project level is not (yet) available, so the firm level is the only available level for an efficiency study of the construction industry at the micro level in Norway.[2]

It should be noted that the firm level should not necessarily be seen as a higher aggregation than the project level. It is not unusual that a project in this industry is larger than any of the participating firms, and quite often a large project can span two or three accounting years.

The layout of the rest of the paper is according to the following plan. Section 2 gives a brief overview of the methods used in this paper. The main ideas are explained, notation is introduced, and the most central references are listed. In Section 2.4 a new approach is developed, that explains the possible benefits of using weighted regression in a two stage DEA setting. Section 3 deals with how the data used in this paper was collected and processed. Selection of the scale specification in the DEA model is the topic of Section 4. In Section 5 results of the DEA efficiency calculations are reported, and the effects of correcting the efficiency scores for bias is shown. Some interesting hypothesis that might explain some of the differences in the firms' efficiency scores are investigated in Section 6. Section 7 rounds off the paper with a summary.

2. The methods

The efficiency scores in this paper are calculated with DEA and then bias corrected with bootstrapping. The model selection is also done with the help of bootstrapping, while the statistical power of the stage two regression is increased by taking advantage of the standard errors of the bias corrected efficiency estimates.

2.1 Data Envelopment Analysis (DEA)

The idea behind DEA is to use the closest possible piecewise linear envelope of the actual data as an estimate of the border of the production possibility area. A more detailed explanation than is given here can be found in e.g. Cooper *et al.* (2000). The efficiency of an observation (often referred to as a "DMU," Decision Making Unit, in the DEA literature) is

[2] Collecting data at the project level is part of the research within "Productivity in Construction."

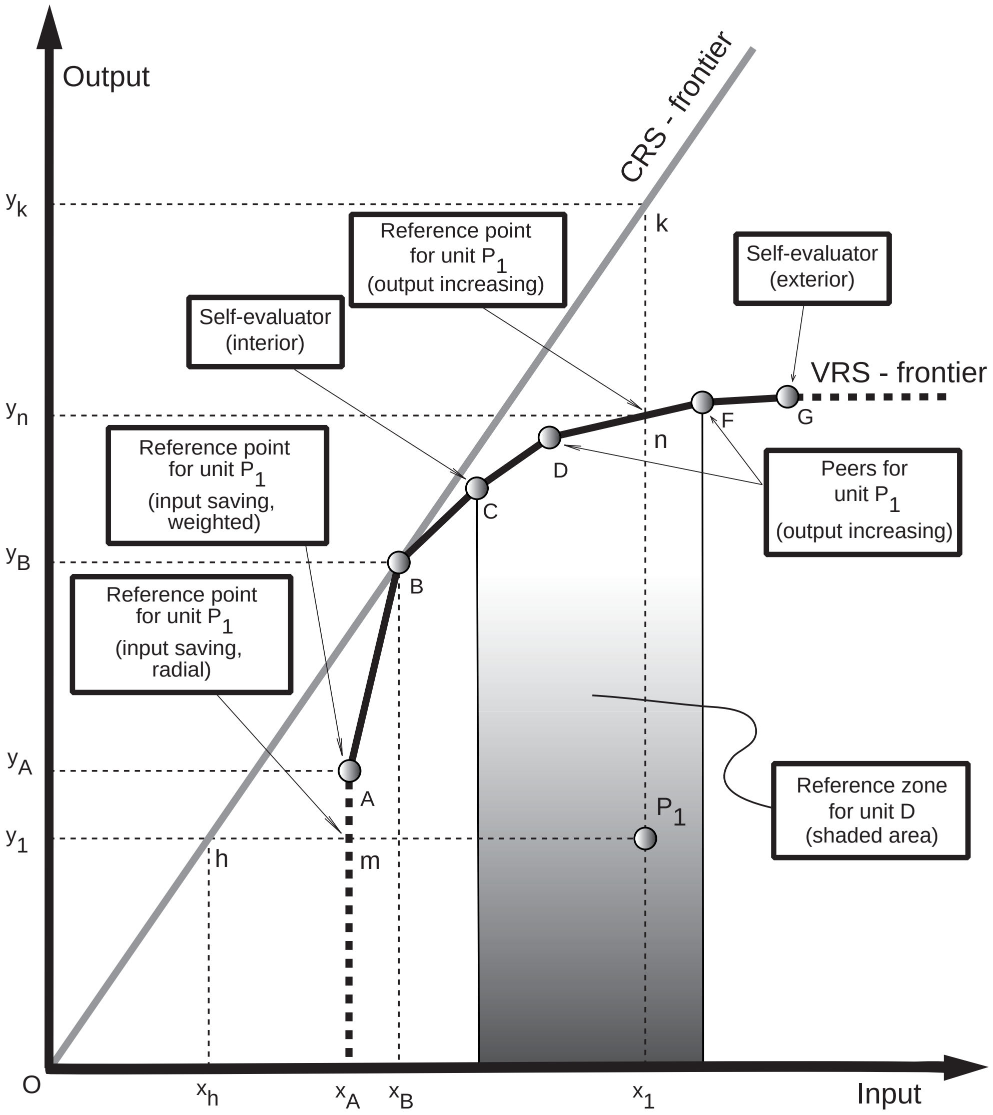
Output
CRS - frontier
y_k
k
Reference point for unit P_1 (output increasing)
Self-evaluator (exterior)
Self-evaluator (interior)
VRS - frontier
y_n
F
G
n
Reference point for unit P_1 (input saving, weighted)
D
Peers for unit P_1 (output increasing)
C
y_B
B
Reference point for unit P_1 (input saving, radial)
y_A
A
Reference zone for unit D (shaded area)
P_1
y_1
h
m
O
x_h
x_A
x_B
x_1
Input

Figure 1: The DEA method illustrated in two dimensions.

calculated as the relative distance to the frontier. The efficiency score is a number between 0 and 1, and the units positioned on the frontier are assigned the efficiency score of 1. Input saving efficiency is a measure of how much it is possible to simultaneously reduce all inputs, while the outputs are at least the same. Banker *et al.* (1984) formalized the axioms that an envelopment should satisfy, and showed that the DEA production possibility set is the smallest set that satisfies the following assumptions (x_1, x_2 are vectors of inputs; y_1, y_2 are vectors of outputs):

1) All observations are possible: If we observe (x_1, y_1), then it is possible to produce y_1 with the use of x_1.
2) Convexity: If (x_1, y_1) and (x_2, y_2) are observed, then $a(x_1, y_1) + (1-a)(x_2, y_2)$ is possible for all *a* in [0,1] (this is true when assuming Variable Return to Scale (VRS). When assuming Constant Returns to Scale (CRS) any positive *a* is allowed).
3) Free disposal: Higher usage of resources always means that it is possible to produce the same or more of products. It is also always possible to produce fewer products with the same amount of resources.

In Figure 1 the most important concepts in DEA are illustrated. A, B, C, D, F, and G are DMUs that in DEA would be calculated as technically efficient when we assume VRS technology, while P_1 is technically inefficient. With CRS only the DMU with the highest output/input ratio would be considered technically efficient, because in this case the productivity of all units is compared – independently of the size of the DMUs. In the following I will concentrate on the VRS production frontier in Figure 1. The reason is that some interesting aspects of the DEA method apply to CRS only if we have more than two dimensions. This again is because under CRS in two dimensions all units are compared to the same face (=the part of the efficiency frontier that the inefficient units can be compared to; each linear part of the frontier) in CRS, but with VRS we typically have more than one facet.

The efficiency measures can be set up mathematically as Linear Programming (LP) problems in the following way:[3]

E_1: Input oriented VRS efficiency can be calculated by solving the following LP problem for each DMU. For unit P_1 in Figure 1 this equals x_A / x_1.

$$
\begin{aligned}
& E_{1i} \equiv Min \quad \theta_i \\
& s.t. \\
& \sum_{j \in P} \lambda_{ij} y_{mj} - y_{mi} \geq 0 \quad , m \in M \\
& \theta_i x_{ni} - \sum_{j \in P} \lambda_{ij} x_{nj} \geq 0 \quad , \ n \in N \\
& \lambda_{ij} \geq 0 \\
& \sum_{j \in P} \lambda_{ij} = 1
\end{aligned}
\tag{1}
$$

[3] See Førsund and Hjalmarsson (1979) where these measures are defined in the general case, independently of the choice of frontier estimation methodology.

The reference point for DMU$_i$ is $(\sum_{j\in P}\lambda_{ij}x_{nj},\sum_{j\in P}\lambda_{ij}y_{mj})$. DMU A in Figure 1 is the input saving reference point for DMU P$_1$. Point "m" is the radial projection point on the VRS frontier, and does not take advantage of the possibility to increase output in addition to the reduction of input.

E$_2$: Output oriented VRS efficiency can be calculated for each DMU by solving the following LP problem. For unit P$_1$ in Figure 1 this equals y$_1$/y$_n$.

$$\begin{aligned}
&1/E_{2i} \equiv Max \quad \phi_i \\
&s.t. \\
&\phi_i y_{mi} - \sum_{j\in P}\lambda_{ij}y_{mj} \le 0 \quad ,m \in M \\
&\sum_{j\in P}\lambda_{ij}x_{nj} - x_{ni} \le 0 \quad , \ n \in N \\
&\lambda_{ij} \ge 0 \\
&\sum_{j\in P}\lambda_{ij} = 1
\end{aligned} \tag{2}$$

E$_3$: Efficiency assuming CRS can be calculated based on either (1) or (2) if we remove the constraint in the last line which demands the sum of the weights to equal one. For unit P$_1$ in Figure 1 this equals either x$_{1/}$ x$_h$ or y$_1$/y$_k$; input and output orientation both return the same number when CRS is assumed.

E$_4$: Input reducing scale efficiency equals E$_3$ / E$_1$.
E$_5$: Output increasing scale efficiency equals E$_3$ / E$_2$.

2.2 Estimating sampling bias using bootstrapping

It is well known that empirical estimations of the DEA models defined in the formulas above are influenced by sampling bias (Simar and Wilson, 1998). The reason is that the DEA estimate of the production frontier is based on a convex combination of best practice observations. If we had sampled all possible DMUs generated by the same underlying Data Generating Process (DGP), we would expect to get a new production possibility area that is strictly outside the DEA estimate. The sampling bias for a given DMU can be expected to be higher the lower the number of other observations in the sample. The DEA frontier estimate is based on the best *observed* practice. But this is a biased estimate of the best *possible* practice in any real world (finite sample) situation.

The following DGP is assumed (Simar and Wilson, 1998): observations are randomly drawn from the true production possibility area. There is a strictly positive probability of drawing observations close to all parts of the true production frontier, and the DEA assumptions (no measurement error, convexity, free disposability) hold. A homogenous efficiency distribution[4] is assumed in the following, but this can be relaxed with a more complicated DEA bootstrap methodology (Simar and Wilson, 2000).

Banker (1993) proved that as the number of draws goes towards infinity, the distance between the DEA estimate and the true efficiency score goes towards zero.[5] In other words, the DEA estimator is consistent. But it is biased in finite samples. The reason is that there is zero probability that a finite number of samples will span the entire outer edge of a continuous production possibility area. The true efficiency of a DMU is the relative radial distance from the DMU to the true production frontier. The DEA estimated efficiency of the same DMU is the relative radial distance from the DMU to the estimated production frontier. The difference between these two distances is the sampling bias. One thing we know is that it is strictly positive, in the sense that the DEA estimated efficiency is higher than the true efficiency.

Simar and Wilson (1998) showed how to estimate the sampling bias in DEA with a method referred to as “bootstrapping” (Efron, 1979). Bootstrapping is in general a way of testing the reliability of the dataset, and works by creating pseudoreplicate datasets using resampling. Bias correction in DEA using bootstrapping is based on the following assumption:

$$(E_1 – E_1*) \qquad \sim(approx.) \qquad (E_1* - E_1**), \tag{3}$$

where E_1 is the true unknown (input oriented VRS) efficiency, E_1* is the original DEA efficiency estimate, and E_1** is the bootstrapped efficiency estimate.

We can not directly calculate the left hand side of equation (3) since the true production frontier is unknown. However, it can be approximated by running computer simulations based on the right hand side of the same equation. This is possible since both the

[4] An efficiency distribution is denoted as homogenous when it is independent of input mix, output mix, and scale.

[5] Banker’s paper showed this formally in the single-output multiple-input case. This has later been generalized to the more general multiple-inputs multiple-outputs (Kneip et al., 2003).

DEA efficiency scores and the linear programs that created them are known. The homogenous bootstrap used in this paper can be calculated with the following algorithm (inspired by Simar and Wilson, 1998):

a) Use the original dataset and calculate the DEA efficiency scores.

b) Create a Kernel Density Estimate[6] (KDE) of the efficiency scores from (a).

c) Move all the DMUs to their comparison point on the frontier.

d) Create a pseudo-dataset by dividing the input values from (c) with values obtained by drawing randomly from the KDE in (b) with *reflection* (Silverman, 1986).

e) Calculate a new series of efficiency scores on the pseudo-dataset in (d).

f) Repeat (d)-(e) a large number of times (2000 is recommended by Simar and Wilson, 1998).

For E_1** in (5) the average value of the efficiency scores in (f) is used.

KDE is used to smooth the empirical distribution of the original efficiency scores (bootstrapping without smoothing is referred as “naïve bootstrapping”). Reflection is used to deal with the boundary condition that is problematic for nonparametric density estimation. The reason is that the KDE smoothing typically results in part of the smoothed distribution densities at values greater than 1.

Denote the difference between E_1* and E_1** with Bias*. Based on this we can create a bias corrected efficiency estimate. Silverman (1993) gives a warning against using bias correction carelessly. The danger is that the bias corrected estimator might have a substantially greater standard error than the original estimator. The result is that we might end up with a new estimator that is unbiased, but at the same time “more wrong on average” than the original biased estimator (larger MRSE). See Simar and Wilson (1998) for a more detailed description.

One difficulty with the algorithm above is that the kernel density estimation requires two parameters: the kernel function (i.e. Gaussian) and the bandwidth parameter (determining the length of the tails of the kernel function). In practice, the choice of the kernel is not nearly as important as the choice of the bandwidth. The theoretical background of this observation is

[6] Kernel Density Estimation is a way getting a smoother estimate of an empirical distribution (when the true shape of the distribution is unknown). See Silverman (1986) for details.

that kernel functions can be rescaled such that the difference between two kernel density estimates using two different kernels is almost negligible (Marron and Nolan, 1988).

In the kernel literature it is documented that the standard formulas for choosing bandwidths pick too large of a bandwidth if the distribution is multi-modal or highly skewed (Silverman, 1986). Since the latter may well be the case for efficiency distributions we avoid using the normal reference rule. Applying leave-one-out cross validation[7] would probably have been the best alternative (Efron and Tibshirani, 1993). The main reason is that selecting bandwidth based on a predetermined mathematical formula would be less subjective.[8] Bandwidth in this paper is selected using visual inspection of the kernel density estimate. This is done using an interactive tool[9] created in Object Pascal for visually inspecting the effects of different bandwidths. The bandwidth selected was 1.0. The effects of choosing other bandwidths (0.5 and 1.5) have also been examined. It made quite a large difference for a low number of units, but the difference in the overall distribution of the efficiency scores was relatively small.

2.3 Testing the scale specification using bootstrapping

As pointed out in Simar and Wilson (1998) the question of whether the production possibility set exhibits CRS has not only economical but also statistical importance. If the true technology is globally CRS then both E_3^* and E_1^* are consistent estimators of the true E_3, but E_1^*might be less efficient than E_3^* in a statistical sense due to slower convergence.

Simar and Wilson (1998) suggest several tests of scale specification using a bootstrapped test. One alternative is the mean of the ratios (with their notation): [10]

$$\hat{S}_1^{crs} = n^{-1}\sum\nolimits_{i=1}^{n} \hat{D}_i^{crs}\left(x_i, y_i\right) / \hat{D}_i^{vrs}\left(x_i, y_i\right) \tag{4a}$$

Using the Førsund and Hjalmarsson (1979) notation (Section 2.1 in this paper):

[7] Leave-one-out cross validation is a technique to investigate the probability that a certain observation was drawn from the same underlying population as the rest of the sample. See Silverman (1986) for details.

[8] As far as is known, there is not any generally available tool for bandwidth selection based on cross validation.

[9] This is a computer program (for the Win32 platform, or Linux under "Wine") developed by Dag Fjeld Edvardsen.

[10] It might be worth mentioning that the notation in Simar and Wilson's (1998) formula 4.1 is a bit unclear since that paper uses a "hat" symbol on top of both the nominator and on the denominator. However, if one reads their paper carefully, one will see in the text that they clearly state that it is the ratio that should be estimated for each of the iterations, not the nominator and the denominator in separate iterations. This is to ensure simultaneous estimation.

$$\hat{S}_1^{crs} = n^{-1}\sum\nolimits_{i=1}^{n} \hat{E}_3^i(x_i, y_i) / \sum\nolimits_{i=1}^{n} \hat{E}_1^i(x_i, y_i) = n^{-1}\sum\nolimits_{i=1}^{n} \hat{E}_4^i(x_i, y_i) \qquad (4b)$$

The question is whether the average scale efficiency we observed (using uncorrected DEA; E_4^*) could have been generated by a CRS technology. An attempt to answer this is made by running a bootstrap simulation where we assume that the true technology is CRS. In each of the iterations we record the average value of E_4. If the average E_4^* that we originally calculated using DEA is outside the given density range, e.g. 95%, then we choose to discard the H_0 that "The true technology exhibits CRS" and use VRS instead.

In addition to the test above, Simar and Wilson (1988) suggest several other tests; among these were the ratio of the means:

$$\hat{S}_2^{crs} = \sum\nolimits_{i=1}^{n} \hat{E}_3^i(x_i, y_i) / \sum\nolimits_{i=1}^{n} \hat{E}_1^i(x_i, y_i) \qquad (5)$$

They end up recommending the ratio of the means (S_2) since it performs best in the Monte Carlo tests. But the mean of the ratios (S_1) performs almost as well, and has an intuitive geometric interpretation. In this paper both S_1 and S_2 will be calculated in the scale specification test.

2.4 Weighted regression in stage two

In the empirical DEA literature it is common to use a "two stage" approach[11] when efficiency estimates are to be both measured and "explained." The first stage refers to the DEA calculation of efficiency scores, based on the data on inputs and outputs. In the second stage it is investigated whether the efficiency scores from stage one are empirically correlated with other variables we believe may "explain" the efficiency scores. The variables used in stage two are typically environmental or managerial variables (both discretionary and non-discretionary variables are commonly used). This possible "empirical correlation" is investigated using multivariate regression models with the efficiency score on the left side of the equation, though other approaches can also be used.

It is often argued that one should do the estimation of both the efficiency explanatory variables and the efficiency itself in the same stage. This argument for improving statistical

[11] See Førsund and Sarafoglou (2002) for a historical account of the origin of the two stage approach.

efficiency is frequently put forward in “standard econometrics.” However, there might be situations where that is not possible or desirable. If it can be assumed that the explanatory variables affect the production in a different way than the regular inputs, i.e. that the explanatory variables do not influence the rate of substitution between the latter, then one might not lose statistical efficiency by using the two stage approach. The explanatory variables might have the character of general shift factors.

Another reason for choosing a two-stage approach is if the explanatory variables are too “rough” for DEA. The assumptions behind the DEA model do not allow measurement error, even in those cases where the measurement errors can be assumed to be symmetrically distributed. A second stage regression model will in such a situation be more robust than DEA. In addition there are more tools available for doing diagnostics and for correcting possible problems. A reason that has been mentioned when these discussions arise is that the explanatory variables are non-discretionary. However, this is not a good motivation to avoid a single stage approach. Several of the generally available DEA software packages allow making one or more of the included variables “fixed” (non-discretionary). An additional argument is that we might not know if the variable is an input or and output (this has been addressed in Simar and Wilson, 2001). Lastly, it might be difficult to include the variable in the single-stage DEA calculation if we have reasons to believe that the relationship between this variable and the efficiency score is not monotonic. It might be partly dealt with either by transformation of data, or possibly by relaxing the assumption of free disposability (allowing for congestion).

Given that a two stage approach is chosen it will be more efficient, statistically speaking, to bring with us the inverse of the standard errors over to stage two. See for instance Carrol and Ruppert (1988) for a more detailed description of using weighted regression from an econometric perspective.

The motivation for using weighted regression is that we have different degrees of certainty when it comes to the precision of the estimation of the efficiency scores in stage 1. For this reason we want to put a larger weight on the more precise observations when we fit the regression hyperplane to the data. This means that there is a greater penalty if the hyperplane is far away from observations with a high weight than for those with a low weight.

When we do the bootstrapped bias correction as described in Section 2.2 we get not only bias corrected point estimates, but also standard errors. These standard errors are (given the assumptions) good measures of how certain we are of the value of each of these point estimates. Since higher standard errors mean lower certainty, the inverse of the standard errors will be used as weights in the stage two regression.

Simar and Wilson (2003) describe two possible approaches for how estimation can be done in a statistically consistent way in DEA in a two-stage setting. They suggest using either a single or a double bootstrap approach, and then argue that the latter is preferable since it has a more rapidly declining MRSE of the intercept and the slope in the regression. Comparing these methods (analytically or using Monte Carlo simulations) with using weighted regression and discovering which of them performs best is a task for further research.

One of the motivations behind this paper is to investigate causes for the efficiency differences among firms in the Norwegian building industry. As described later in this paper (Section 2.4), weighted regression in a two stage setting will be used, and the reason is to reduce the influence of the bias corrected efficiency scores with a large estimated standard error. The view of the current paper is that an unbiased estimator might be useful even if it has an estimated MRSE larger than the original biased one. One example is to use it in a second stage regression (with weights based on the inverse of the standard errors), which is done in Section 6.3.

3. The data

In 2001 the Norwegian building industry consisted of about 34 500 enterprises, and employed about 132 500 persons (about 10 percent of the Norwegian labour force). From 2000 to 2001 there was an 8.7% growth in turnover and 7.4% growth in employee compensation. The efficiency calculations in this paper must be seen in the light of the fact that the industry experienced strong growth in the year under investigation.

The primary data on the building enterprises is collected on a yearly basis by Statistics Norway. All the firms in the dataset used in this paper have a NACE code of 45.211. This means that at least 50% of their production value is in the category “construction of buildings.” The sample collected by Statistics Norway consists of all enterprices with more than 100 employees, and a sample of the smaller enterprices. The sample contains at least 30% of the total emplyment in the NACE 45.211 subgroup.

Based on data for each building enterprise we have created a cross section database on production and resource usage for the most recent year available (2001). A rather extensive set of input and output data were available based on annual company accounts and the structural survey conducted by Statistics Norway. After extensive discussions with Statistics

Norway and sector experts the input-output specification was selected. Output is measured as value split on three different categories: Residential buildings, Non-Residential Buildings, and Civil Engineering. The three inputs are External Expenditure, Labor, and Real Capital. Details are laid out in Section 3.2.

3.1 Data quality filters

Statistics Norway has several routines for detecting and correcting erroneous data. This should help improve the quality of the data. However, the data collected in the yearly surveys is for general purposes, and the definition of which observations we believe to have good enough quality depends on what they are to be used for. Productivity measurement with frontier models is especially sensitive to outliers. When we use the DEA model we formally assume that there are no measurement errors in the data we feed into the efficiency estimation model.

However, it is important to avoid shaping the results to confirm *a priori* suspicions. This is especially important in a frontier setting, because the frontier-defining units are by definition outliers. But experience with empirical DEA applications strongly suggests that not cleaning the data for suspicious units can lead to very questionable and sometimes absurd results. Very influential units should be checked extra carefully, since errors in these DMUs can strongly influence the efficiency estimate for a large number of other DMUs.[12]

It was required that all the observations used in the DEA model should be able to meet the following three requirements. (1): At least 90% of the production (measured in total value) has to be from the construction industry.[13] 15.8% of the companies did not meet this requirement. (2): All three inputs must be greater than zero. 23.6% of the companies did not meet this requirement. (3): The observed usage of labor in man-years must be larger than or equal to one.[14] 4.9% of the dataset did not meet this requirement. 3.2% of the companies failed to meet more than one of the three requirements.

After this automatic cleaning, five more units were removed from the dataset after test runs of the DEA model (using a VRS specification). They showed up as strongly influential

[12] In Thorgersen et al. (1996) the “Peer index” was introduced. This measures the influence of each of the peers in the DEA estimation relative to how large a share of the improvement potential (for each of the dimensions) this peer is referencing. The calculation of the Peer index is based on the optimal weights in the DEA calculation. The maximum value is 100%, and can only be attained if this peer refers all potential improvements in this dimension. The Peer index is a useful measure of the influence of each of the peers in the dataset.

[13] The number 90% is ad-hoc, but is selected to make sure that the building firms are homogenous in the sense that none of them are allowed to have a large share of their sales outside the building industry.

(large peer index, see Torgersen et al., 1996) and with high superefficiency[15] (see Andersen and Petersen, 1993). Originally superefficiency was used as a way to rank among the efficient units, but recently it has more often been seen as a way of detecting strange units. Removing strongly influential units if they are radial outliers might be questionable, but is probably the least evil choice.

3.2 Descriptive statistics for the primary dataset

The resource usage of the building entrepreneurs is captured by three inputs (the three first columns in Table 1): *External Expenditure* includes materials, subcontractors, energy, transportation etc. *Labor in Man-Years* is a measure of the labor usage. *Real Capital* is a measure capital service based on the use of production equipment, machines, etc. It is calculated from rental expenditures and depreciation. The last three columns of Table 1 contain summary statistics on the production of the building entrepreneurs. *Residential* is a measure of the sales value of the residential and recreational buildings. *Non-Residential* is a measure of the sales value of other buildings, such as office buildings and institutional buildings (schools, prisons, hospitals etc). *Civil Engineering* measures the sales value of constructions such as roads, tunnels, harbors, etc.

Because of the data filters all three inputs have strictly positive values, and the lowest value for Labor in Man-years is 1. The lowest value for all the three product variables is 0, but all firms have a strictly positive sum of output values. Construction is clearly the output with the lowest number of strictly positive output values, and is also the variable with the largest CV-number. Concerning the size distribution, 39% of the firms use less than 10 man years,

Table 1: Descriptive statistics for the primary variables (342 observations after the data cleaning).

	External Exp.	*Labor in Man-years*	*Real Capital*	*Residential*	*Non-Residential*	*Civil Engineering*
Minimum	18.0	1.0	2.0	0.0	0.0	0.0
Maximum	4 083 634.0	2 950.0	23 9847.0	1 597 609.0	2 556 175.0	1 170 239.0
Average	44 968.6	38.3	1970.1	27141.4	32 160.2	5 688.4
St.dev	233 063.7	167.2	13 655.7	113 635.8	147 942.9	66 574.4
Count >0	342	342	342	301	218	31
CV[16]	5.2	4.4	6.9	4.2	4.6	11.7

[14] The reason is that only real production firms are included. Firms not meeting this demand may be pure accounting units, or may be newly started, closing down or in hibernation.

[15] Superefficiency is a measure of the relative radial distance from the origin to the DMU in question, when the frontier is estimated without this DMU included in the dataset. Superefficiency is by construction greater than (or equal to) one. A superefficiency value of 1.2 implies that the DMU is positioned “20% outside” where the frontier would have been without this DMU (in a radial sense).

[16] CV is the Coefficient of Variation. It is defined as the ratio of the standard deviation and the average.

47% use between 10 and 50 man years, 11% use between 50 and 100 man years, while 5% use more than 100 man years . The average firm has close to 41 employees and uses 38 man years.

4. Choosing model specification

In Section 2.3 it is explained how one can use the bootstrapping methodology to help select the scale specification of the DEA model. If the scale efficiency (E_4) in the original DEA model is outside the (95%) one sided lower confidence interval we reject the H_0 that the technology is CRS and apply VRS instead. In Figure 2 the bootstrapped simulations required are plotted in a histogram. If the null hypothesis were true we would expect the observed E_4 from the uncorrected DEA estimation to be located the inside the 95% confidence interval. The observed E_4 is 0.777 and we get a strong rejection of H_0.

The histogram of S_2 is practically identical. In both cases (S_1 and S_2) we get a very solid rejection of the null hypothesis ("The true production technology is globally CRS"). Based on this result we will in the following assume that the technology exhibits variable returns to scale.

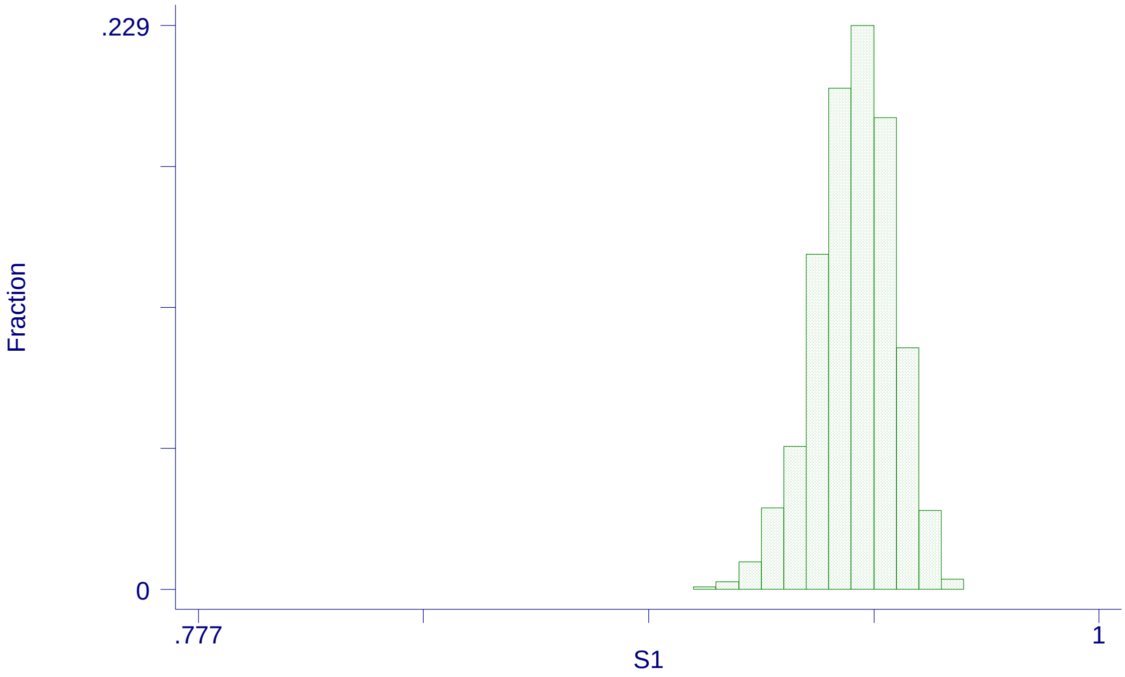

Figure 2: Histogram of the bootstrapped distribution of the average scale efficiency (S_1) assuming CRS.

In this paper the only statistical tools for choosing the correct model are the ones designed for tests of scale specification. A similar set of bootstrapped tests for model specification could also be used for selecting which variables should be included. This line of thought is based on Banker (1993, 1996) and Kittelsen (1993), but it would be better from a statistical perspective to perform these tests using the bootstrap methodology. However, it is important to select the model based on economic theory and the knowledge of the sector we are investigating – not purely on statistical tests.

5. Estimating the efficiency scores

5.1 DEA efficiency scores

The figures showing the uncorrected DEA efficiency scores (E_1 and E_3) will only be commented on briefly since the numbers change greatly when correcting for sampling error using the bootstrap method. However, it is important to point out that almost all of the published DEA papers stop with calculating only the DEA efficiency scores, and do not estimate and correct for sampling bias. It will be shown that this makes a big difference for the interpretation of the results. Refer to Section 2.2 for explanation of bias correction

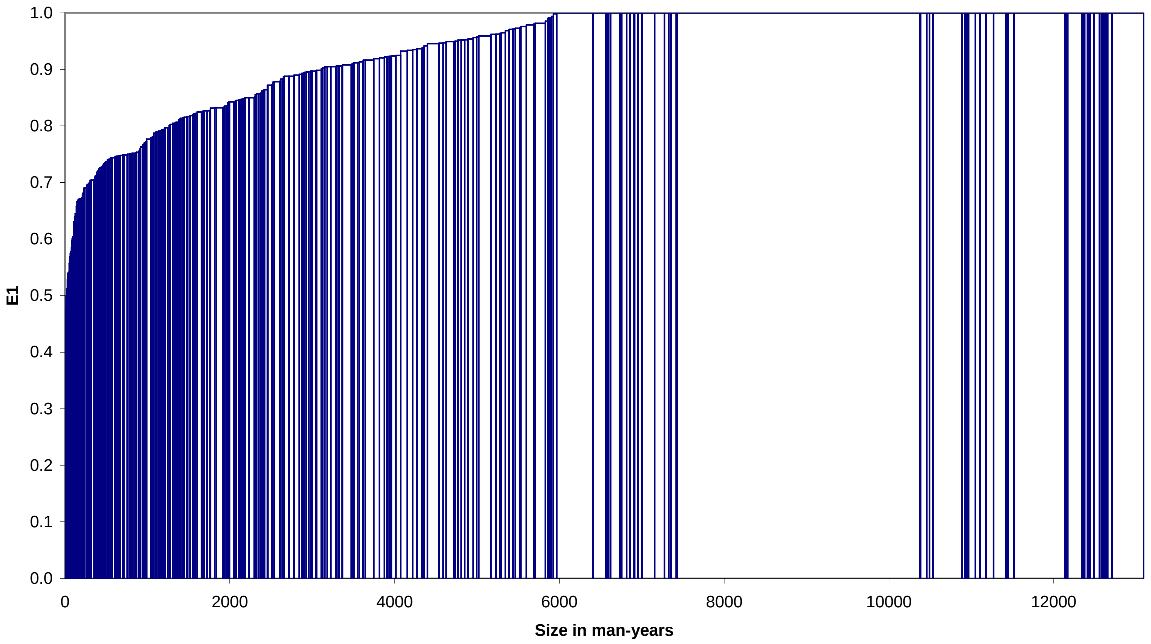

Figure 3: Uncorrected efficiency scores assuming variable returns to scale.

.

Figure 3 shows the uncorrected efficiency scores, assuming VRS, in an *Efficiency diagram*.[17] One interesting feature of Efficiency diagrams is that both the height and the width of the bars can contain information – unlike a bar chart where only the heights of the bars are actively used. This is especially useful when illustrating the results of efficiency analysis. The efficiency of each of the DMUs is shown by the height of the bar, while its economic size (man-years in Fig. 3) is shown by the width of the bars. This means that it is possible to examine whether there are any systematic correlations between the sizes of the units and their efficiencies. Another interesting geometric aspect of these figures is that they are sorted according to increasing efficiency from left to right. The distance from the top of each bar to 1.00 is a measure of that particular DMU's inefficiency, and the width of the bar is a measure of its economic size. For this reason the area above each of the bars is proportional to the economic cost of that DMU not being 100% efficient. This means that there will typically be a "white triangle" above the inefficient units, and that the size of this area is proportional to the economic cost of the total inefficiency in the sample. The software used to construct these graphs is an "add-in" for Microsoft Excel.[18]

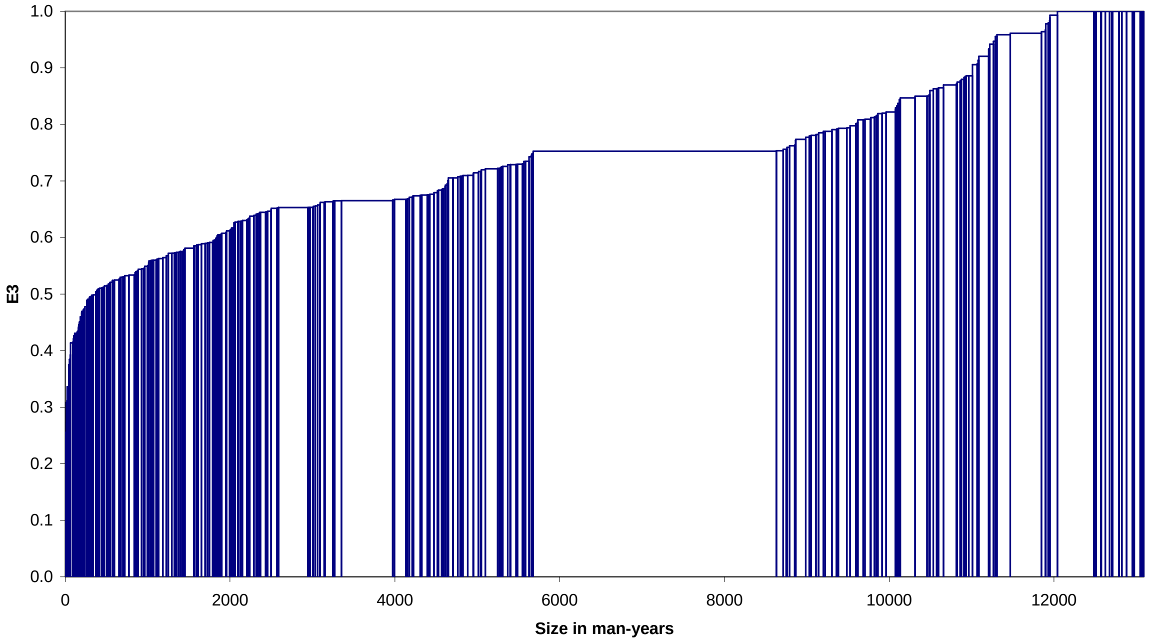

Figure 4: Technical productivity (E_3).

[17] This diagram was first introduced in Førsund and Hjalmarsson (1979), and is based on the input coefficient diagram in Salter (1960).

[18] This add-in (MS Excel under Win32) is authored by Dag Fjeld Edvardsen, and will be made available for free from http://home.broadpark.no/~dfedvard.

When it comes to interpreting Figure 3 there are two aspects that are dominating. The first is that the average efficiency is quite high (83.44% to be exact), and the second is that all of the largest units have been considered fully efficient.

Figure 4 shows efficiency under CRS (E_3). E_3 is still interesting even when we have chosen VRS as the correct scale assumption. The reason for this is that E3 is also a measure of "technical productivity," so it is useful even when we don't believe in it as a measure of efficiency. The efficiency is much lower with E_3, and the difference is quite striking for the largest units. However, when comparing Figure 3 and 4, it is important to remember that sampling error has not been taken account of. Since the difference between the LP formulation for E_1 and E_3 is that the former implies restrictions on the multiplier weight, we have reason to believe that the efficiency measure E_1 is more affected by sampling error than the productivity measure E_3. A rough explanation is that with E_3 all units are potentially compared independent of size, so each DMU has a higher number of units to be compared with.

Figure 5 shows the histogram of the uncorrected efficiency scores from DEA, while Figure 6 shows the same for the bias corrected efficiency scores. It is obvious to the eye that the change of the distribution is dramatic. The most obvious difference is that the strong concentration of fully efficient units at the right of the histogram disappears when we correct

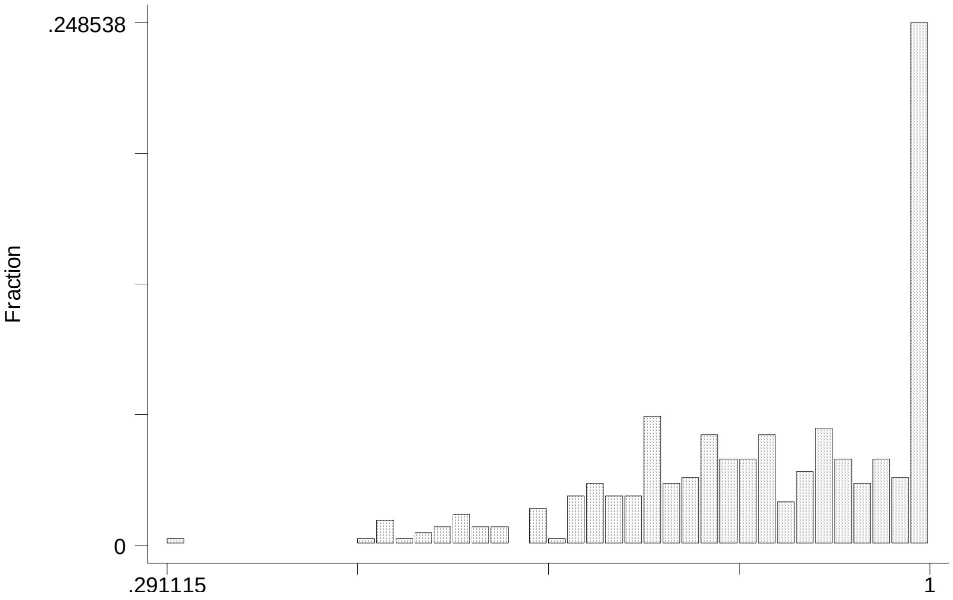

Figure 5: Histogram of uncorrected efficiency scores (E_1).

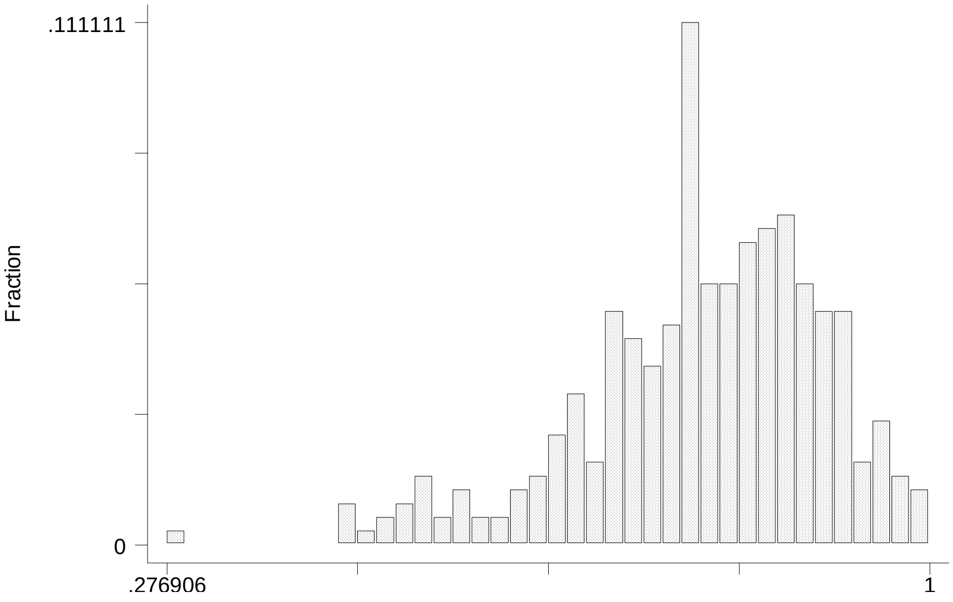

Figure 6: Histogram of bias corrected efficiency scores (E_1).

for the estimated sampling bias. In fact, only three DMUs are assigned unit efficiency score after the bias correction.

Figure 7 contains both the bias corrected and the uncorrected efficiency scores in an Efficiency diagram. The bias corrected values are the lower bars, while the uncorrected values are plotted in the upper curve. Both series are sorted independently of each other. It is obvious to the eye that the estimated inefficiency is much larger with the bias corrected values.

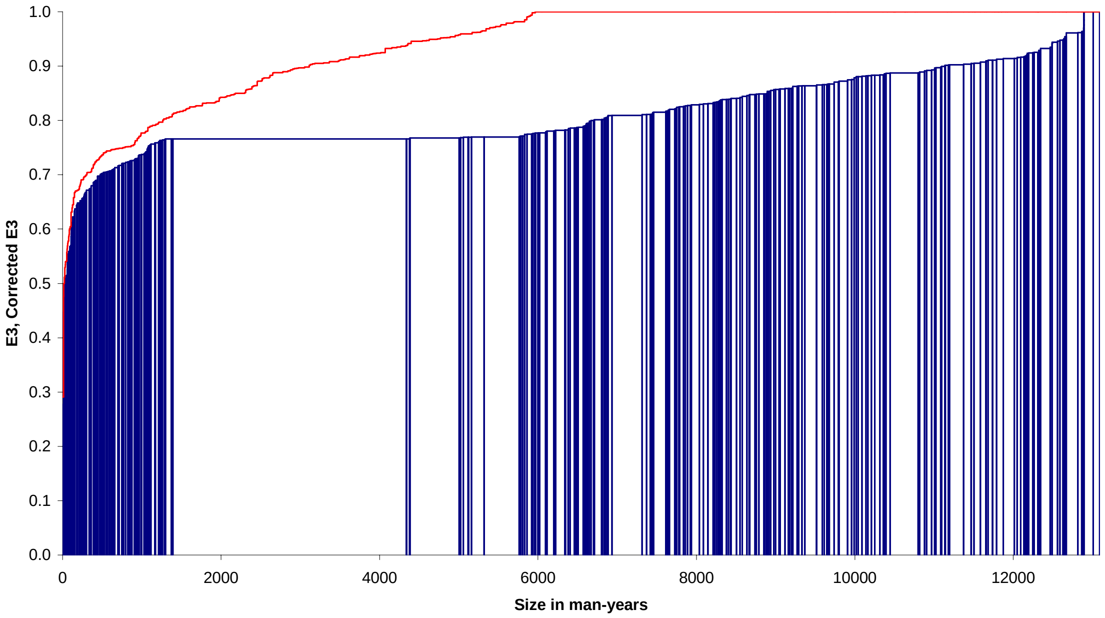

Figure 7: Efficiency diagram of bias corrected and uncorrected efficiency scores (E_1), sorted separately.

Figure 8 is based on Figure 7, but the bias corrected efficiency scores are sorted pairwise with the uncorrected efficiency scores. This allows us to compare the efficiency score for each individual DMU before and after the bias correction, and also examine whether there is any systematic difference when it comes to how the sampling error influences firms of different sizes.

Inspection of Figure 8 confirms that all of the large construction firms have a large estimated bias. This is often the case, because the sampling error typically hits the large firms harder in a VRS model. The tendency is relatively strong, as shown by the regression on estimated bias versus production volume (and its square) below. Figure 5 is quite instructive because it shows the big difference that the bias correction does with the very large units. This strongly suggests that analyzing scale economies without checking for sampling bias in DEA can give misleading results. In addition, since the large firms very often contribute a large share of the production and resource usage of an industry, measures of efficiency at the aggregated level will tend to be more distorted.

The same problem is present for the smallest firms, but this is much more difficult to point out in Figure 8 since the width of the bars are proportional to the size of the firm. An OLS regression between the estimated bias and the size of the firm (measured by the sum of sales and its square) obtains statistical significance for both parameter estimates (and the intercept), and the R-squared is 0.1755.

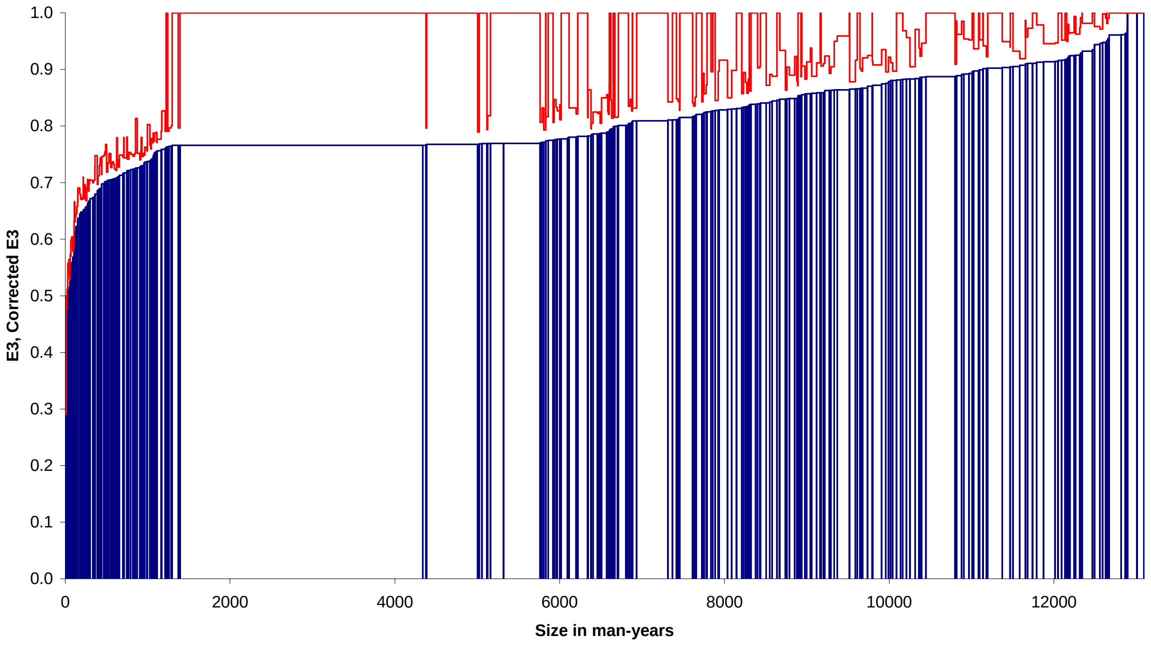

Figure 8: Efficiency diagram of bias corrected and uncorrected efficiency scores (E_1), pairwise sorted.

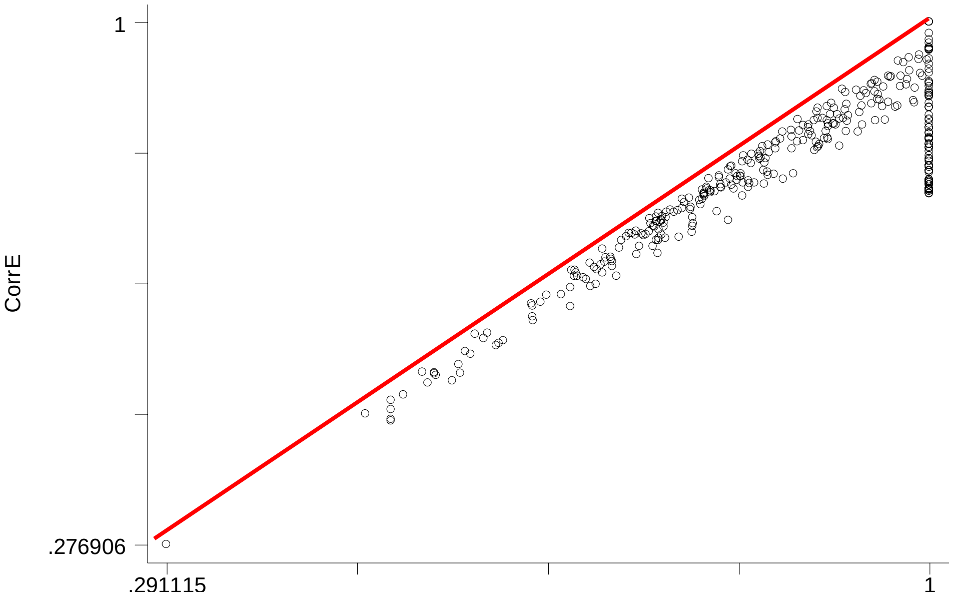

Figure 9: Scatter diagram of bias corrected and uncorrected efficiency scores (E_1).

The suggestion that bias tends to be correlated with scale is a little alarming. One reason is that we might get distortions when we investigate the economics of scale. Another reason is that it might not be unusual that the explanatory variables are correlated with the size of the DMU. A statistically significant finding of the correlation between uncorrected DEA and the explanatory variable might be the result, and misinterpretations are in these cases easily made. The effect might be strongest in VRS, where the effects of scale are supposedly removed from the equation. Practitioners that have carried out regression analysis on the uncorrected VRS efficiency scores without checking for the effects of scale (because it was supposedly already taken care of in VRS) might want to revisit old results and see if this happened in their case.

Figure 9 shows almost the same information as Figure 8, but in a scatter diagram. The information about the size of the units is removed, but on the other hand it is easier to compare the individual changes with and without bias correction. The units with an uncorrected efficiency score of 1.00 are the ones that are changed the most by the bias correction.

Table 2: Original efficiency scores and bias corrected efficiency scores.

Orig.Range	**Avg CorrE**	**Obs.**
1	0.855	77
0.9-1.0	0.891	69
0.8-0.9	0.809	76
0.7-0.8	0.723	71
0.6-0.7	0.634	28
0.5-0.6	0.513	19
0.4-0.5	0.457	1
0.3-0.4	--	0
0.2-0.3	0.277	1
	Sum Obs.	342

Table 2 shows the original e-scores (not corrected for sampling bias) in its left column. In the middle column are the average bias corrected efficiency scores for the same observations. It is interesting to note that the average bias corrected e-score for the units with unit uncorrected e-score is lower than the e-scores for the group with uncorrected e-score in the range between 0.9 and 1.0. This is a reminder of the large sampling bias associated with the DMUs that were assumed to be fully efficient in the uncorrected DEA calculations.

The units we want to learn from are the ones with the highest corrected efficiency scores and the smallest confidence interval. When choosing between combinations of these two attractive aspects it would probably be wise to focus on the DMUs whose efficiency scores have low standard errors among the DMUs with quite high corrected e-scores. This is done in contrast to being totally focused on choosing the DMU(s) with the very highest corrected efficiency scores.

Table 3 displays the descriptive statistics for the VRS (E_1) and CRS (E_3) efficiency scores, with and without bias correction. Before examining the numbers, one might expect the difference between the uncorrected and corrected average efficiency score to be largest in the VRS case. The reason is that the problem with sampling error might be expected to be largest in the VRS case, since the LP formulation (2) for the DEA problem requires the sum of reference weights to equal one. This is not the case in the CRS model, and for this reason

Table 3: Descriptive statistics for VRS and CRS efficiency scores, with and without bias correction.

	Obs	Average	Stdev	Min.	Max
E_1 uncorrected	342	0.848	0.136	0.291	1
E_3 uncorrected	342	0.687	0.167	0.255	1
E_1 biascorrected	342	0.785	0.116	0.277	1
E_3 biascorrected	342	0.615	0.137	0.235	1

there is a higher number of potential units that a given DMU can be compared with under the CRS assumption. But the difference between the uncorrected and bias corrected average efficiency score is slightly larger under CRS than under VRS in Table 3.

R^2 for a regression between the uncorrected and the corrected efficiency scores is 0.801 if we allow for an endogenous constant term in the fitted regression equation. The fit is quite good, but as mentioned the difference is large for a high number of units.

There is a tendency (as seen in Figure 10) that the lower the original efficiency score, the lower the estimated sampling bias. There might be several reasons for this, but one possible explanation is that the more centrally positioned a DMU is in the dataset (with regards to size and input/output mix), typically the higher the number of units it will be compared with.

In other words, we expect a unit with a low observed uncorrected DEA efficiency score to be closer to its real value than one with a high observed efficiency score. The reason is that an observed uncorrected efficiency score is expected to be correlated with centrality in the dataset. The bias can be expected to be the largest for the units with the highest scores in the uncorrected DEA model, and the lower the uncorrected efficiency score the lower the expected bias. But the decrease in bias seems to be slower the lower the efficiency score of the DMU under investigation (a smooth fitted parametric function would have a positive first and second derivate.

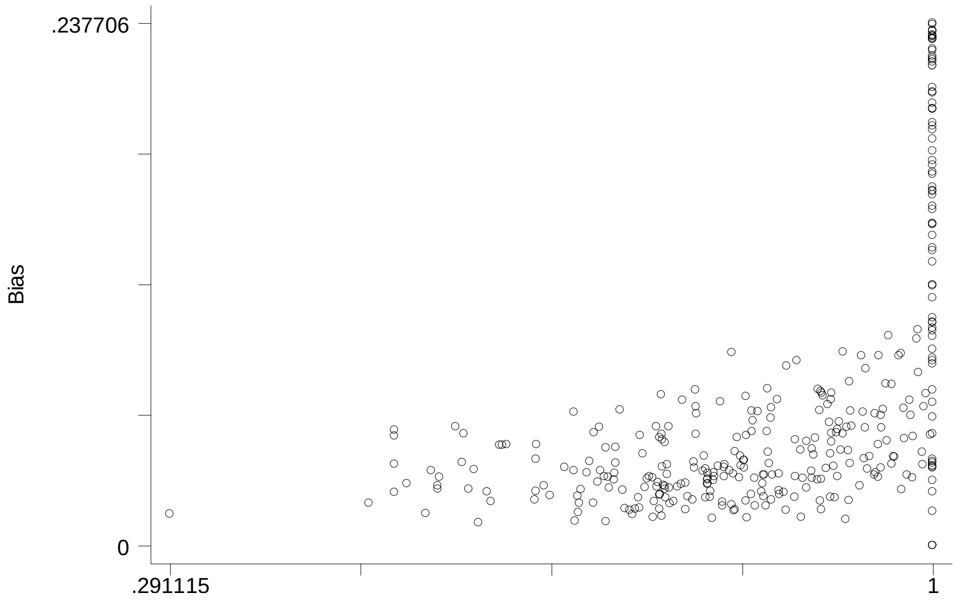

Figure 10: Scatter diagram of estimated bias vs the original uncorrected efficiency scores (E_1).

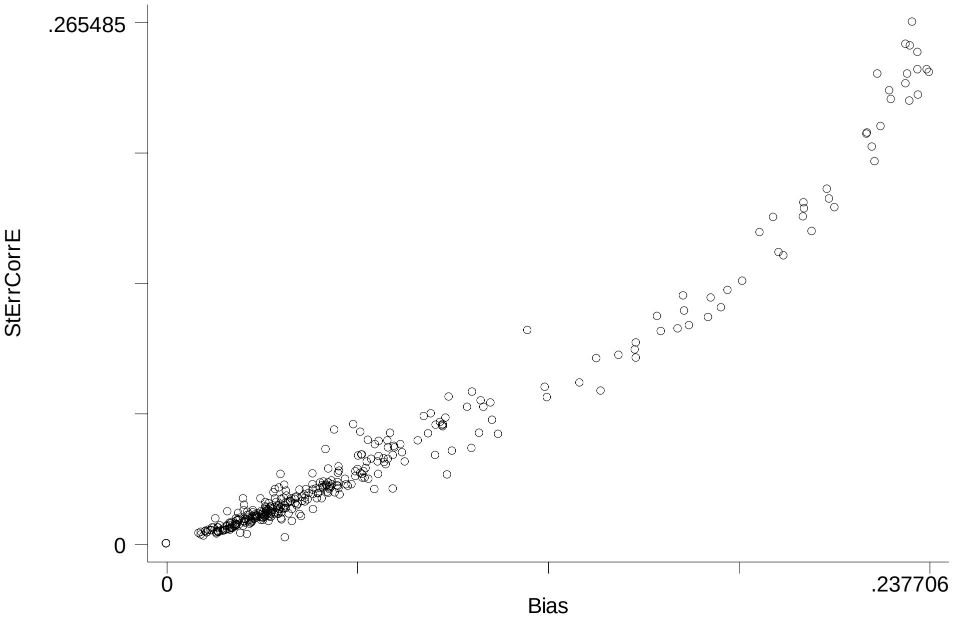

Figure 11: Scatter diagram of bias versus standard error of the corrected efficiency scores.

This is closely related to the problem with the curse of dimensionality and the speed of convergence of the DEA model relative to sample size. It might be that the empirically strong relation discovered in the dataset can be investigated in light of a theoretical relationship. Such a formula might show the relation between the true efficiency and the size of the estimated bias. It seems that the lower the true efficiency, the lower the size of the bias (Bias = f(E1), f'>0, f''>0) . However, this task will have to remain for further research.

Figure 11 shows the strong empirical correlation between the estimated bias and the standard error of the bias corrected efficiency estimate. The relation is so strong that one might want to examine if this can be established formally: The higher the estimated bias, the higher one can expect the standard errors to be. A possible explanation is that the high estimated bias occurs for samples that one has little information about. If the area of input/output space is scarcely populated, we have little information about the location of the frontier; especially if the area is outside the center of the sample (we get little help from the convexity assumption and surrounding observations). At the same time, this lack of information is also captured in a large standard error. In other words, there might not be two separate phenomena, but rather two manifestations of the lack of information about the area in the input/output space where the DMU is located. This strong empirical correlation has probably not been shown before in the literature, and an extensive search on the Internet for similar findings turned up with no relevant findings.[19]

As mentioned in Section 2.2 one should be careful when using the bias corrected efficiency estimates without evaluating the standard errors. The reason is that the bias corrected efficiency estimates might get higher MRSE than the original DEA estimates. Only 202 of the 371 DMUs (54.4% of the sample) in the dataset get bias corrected efficiency with lower estimated mean square error than the original DEA efficiency estimate.

6. Efficiency and productivity explained

The efficiency estimation in Section 5 revealed large differences among the firms when it comes to technical efficiency and productivity. In this section, different hypotheses that might "explain" these differences are developed and tested in a stage two setting for empirical relevance. In addition to including the efficiency scores from stage one, it is argued in Section 2.4 that the certainty level for each of the observations should also be taken into account. It should however be noted that the available data is not sufficiently detailed to give a clear indication of why some firms appear to be much more efficient than others. Some hypotheses are associated with statistically significant parameter estimates, but these should mainly be viewed as indications of interesting topics for further research.

6.1 Constructing hypotheses based on existing theory and knowledge of the industry

The hypotheses developed below were generated based on knowledge about the industry and given by the limited availability of data.

a) Wage cost per hour. Higher wages can attract the best workers. In addition, piecework contracts can lead to higher average wages. One might suspect that this factor is closely related to the hypothesis (d) since overtime is also associated with a higher hourly pay. However, a regression with Wage Cost per Hour explained by Hours Worked per Employee gets an R^2 of only 0.03, so there should be no problem with multi-colinearity.

b) Apprentices. This is defined as the number of apprentices relative to the number of employees. The idea is that the most efficient companies have low shares of apprentices. The reason is that we expect the companies with high shares of apprentices to have higher costs

[19] A search was carried out on the Internet search engine Google.com with the terms: DEA bootstrap correlation bias standard error (and confidence interval). It did not return any relevant results.

and lower production since apprentices are under training, which should imply lower productivity of the apprentices and also man-hours used by other employees to offer them guidance. On the other hand, a history of having a high number of apprentices could give good access to high quality human capital in the long run. The hypothesis is nevertheless that the total effect of a high number of apprentices is reduced efficiency.

c) Product Mix. This is measured, using the Herfindahl index, by the quadratic share of the sales of the companies divided between the seven underlying markets. The expectation is that the most diversified firms are more efficient since they have the option of using their resources in the most attractive market depending on short term business cycles. It should be mentioned that the business cycles in the construction industry can change very fast. In addition, there is a possible selection effect since this variable might pick up that the "best" firms get contracts outside of their key markets. Notice that testing the economics of scope using a DEA model can be problematic. The reason is that DEA assumes global convexity. If we find that the most diversified companies are the least efficient, this might be a warning that we have a serious breach of one of the assumptions underlying the DEA model used in stage one.

d) Hours Worked per Employee. The hypothesis is that the firms with high numbers of hours per employee are more efficient. The reasoning is that these firms get more efficient production by the use of overtime. It might be that some employees work better under a certain degree of pressure. It could also be that the best workers choose their employer based on the opportunity to work overtime. An additional possibility is that the repetition effect is positive and that the use of overtime allows for longer repetition sequences.

e) Located in Oslo. Oslo is the capital and the largest city of Norway, and a pressure area. Housing prices are usually higher in the Oslo area, and the way efficiency in this paper may be influenced by this price effect.

It would have been interesting to follow up Albriktsen and Førsund (1991) and examine if the amount paid to subcontractors is correlated with efficiency. However, the quality of the data at hand is too low (a large number of firms have reported a value of zero even when this is not believable).

Table 4: Correlation table for the explanatory variables.

	Wage cost per hour	Share of apprentices	Product mix	Hours worked per employee	Location Oslo
Wage cost per hour	1				
Share of apprentices	**-0.2188***	1			
Product mix	0.0197	**-0.1446***	1		
Hours worked per employee	**-0.1546***	0.0208	0.0842	1	
Location Oslo	**0.2242***	-0.1	-0.0298	0.068	1

Table 4 shows the correlation table for the explanatory variables. The pairwise statistically significant correlations (at the 5% sig. level) are marked with a asterisk (and a bold font style), and only those will be commented on. Wage Cost per Hour is negatively related to Share of Apprentices, Hours Worked per Employee and Location Oslo. In addition Share of Apprentices is negatively related to Product Mix. That Wage cost is negatively correlated with Share of apprentices and Location Oslo is not surprising since the pay to apprentices is less than for trained labor, and the wages in Oslo is known to be higher than in other parts of Norway. It is surprising that Wage Cost per Hour is negatively correlated with Hours Worked per Employee, but it might be that the employees with low hourly salaries choose to compensate by working extra hours. It might be that the substitution effect, known from Labor Economics, dominates. It is difficult to explain why Share of Apprentices is negatively correlated with Product Mix.

6.2 Do the suggested hypotheses have empirical relevance?

The regression models used below are weighted least squares and OLS (for comparison). The weights in the weighted regression are the inverse of the squares of the estimated standard errors from the bootstrap simulations. The motivation is to put low weight on an observation when there is low certainty of its real value. The regression calculation was carried out in Stata7. This statistics package (and many others) has built in support for assigning *a priori* (in a sense) known weights to the observations.

Table 5 shows the results from an ordinary OLS regression (left part of the table) and a weighted least squares regression. A truncated regression was also computed (with right truncation at 1) but the results were as good as identical to those laid out in Table 4. The coefficients with statistically significant parameter estimates (at the 5% significance level) are

Table 5: Weighted and unweighted regression.

Explanatory variables	**Unweighted (R^2=0.14)**			**Weighted (R^2=0.26)**		
	Coef.	t	P>|t|	Coef.	t	P>|t|
Wagecost per hour (a)	**0.0012**	**6.69**	**0**	**0.0019**	**9.14**	**0**
Share of apprentices (b)	-0.0802	-1.15	0.25	**-0.1487**	**-2.18**	**0.03**
Product mix (c)	**-0.0478**	**-2.04**	**0.042**	**-0.0880**	**-3.83**	**0**
Hours worked per employee (d)	**0.0001**	**2.64**	**0.009**	**0.0001**	**4.02**	**0**
Oslo (e)	-0.0106	-0.49	0.622	0.0093	0.32	0.752
Intercept	**0.4349**	**5.91**	**0**	**0.2456**	**3.1**	**0.002**

highlighted with bold fonts. A higher number of the parameter estimates are statistically significant in the weighted regression compared to the OLS regression. The p-values are also lower in the weighted regression, with the exception of the Intercept estimate which has a slightly lower p-value in the OLS regression.

If we believe in the statistically significant parameters from the weighted regression i(Table 5), then the most efficient construction companies are characterized by:

- High average wage per hour
- Low numbers of apprentices
- Low concentrations in product mix
- High numbers of hours worked per employee

Not statistically significant:

- Located in Oslo

Earlier in this paper it has been shown (Fig. 4) that the bias correction and the standard error of the bias corrected efficiency score have a strong and positive correlation. Remember that the inverse of the latter were used as weights in the main regression model. The implication is empirically that low weights are put on the units that have gotten a strong bias correction (because they very often get large confidence intervals). It was noted above that the units with the highest bias correction tend to be found among the units with efficiency scores equal to 1 from the uncorrected DEA calculations.

Many applied DEA papers have used tobit regression in a second stage. The reason is probably that the authors have observed a concentration of DMUs with uncorrected efficiency scores of 100%, and that they based on this have thought of the DEA efficiency scores as being truncated. This is wrong since (as seen in the LP formulation) they are not truncated[20] -- they are serially correlated. Simar and Wilson (2003) shows that using a tobit regression in stage two is both theoretically and empirically (using Monte Carlo simulations) wrong.

6.3 Productivity and scale

Earlier in this paper a bootstrapped scale specification test rejected the null hypothesis that the correct model specification was CRS. However, even when we choose to believe that the true production function exhibits VRS we can find use for the CRS measure, and interpret it as productivity. This can be used as a measure of to what degree the sector has an efficient structure.

In Figure 12 the maximal value plotted on the horizontal axis is 1,200,000. The intention is to zoom in on the range where there seems to be most interesting systematic tendencies in the simultaneous distribution of average productivity and scale. It seems that the average productivity of the construction firms increases until the size is about 100 millions (NOK). There does not seem to be any systematic change after this value is reached. However, there are not many construction firms with production values much higher than 100 millions NOK in the sample, nor in the population of all Norwegian construction firms.

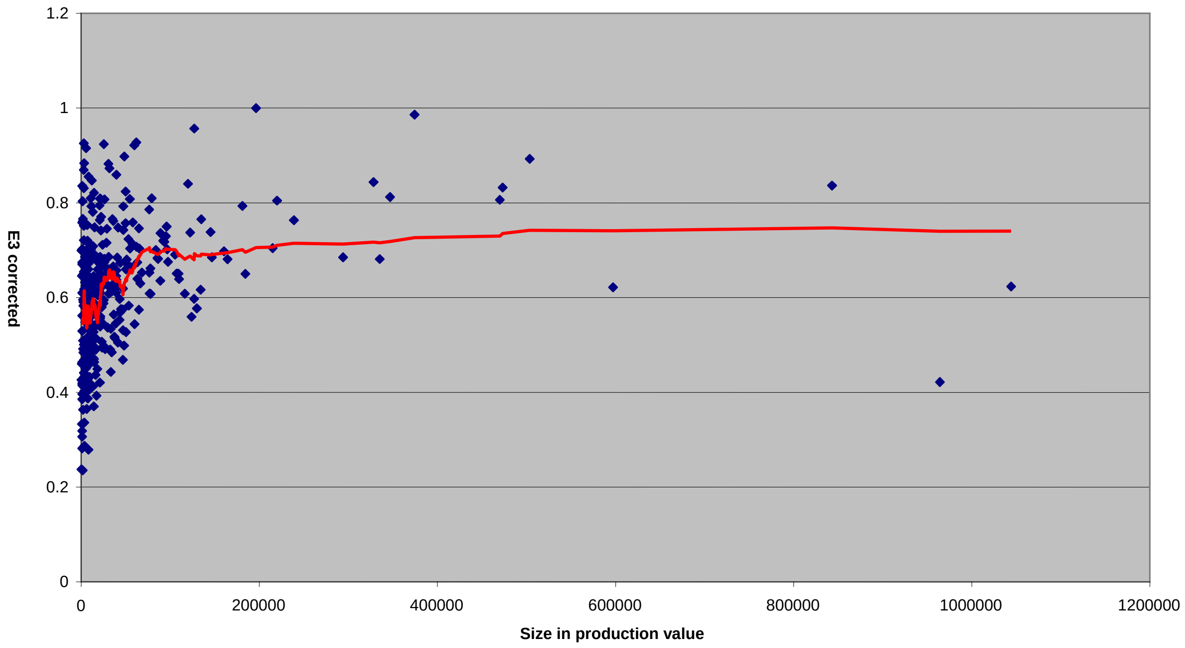

Figure 12: Scale chart showing Production Value and E3 (range 0-1’200’000).

[20] The efficiency scores can never be above 1. The reason for the concentration at 1 is that the efficiency scores of each unit depend on the input-output vectors of the other units (leading to serial correlation in the best-practice calculation). They are not truncated at 1 as such.

7. Conclusions

This paper concerns using DEA to investigate the efficiency of Norwegian building firms. Large differences in the efficiency and productivity scores were discovered. One important lesson that can be learned from this application is the danger of taking the efficiency scores from uncorrected DEA calculations at face value. If one decided to learn from a few DMUs based on their uncorrected efficiency scores, one might get into trouble. It is not unreasonable to think that similar things have happened in the last few years as DEA has been embraced by a very large number of practitioners (researchers and consultants).

It would be interesting if the large number of empirical DEA papers were recalculated using the bootstrap methodology. Anecdotal observations indicate that very few practitioners use bootstrapping. The reason for this might be that bootstrapping is not yet available in the standard DEA software packages.

Based on a scale specification test, a variable returns to scale specification was selected. A scale chart indicated that firms with total production values lower than 100 mill. NOK might be operating at a suboptimal scale level.

The differences in the efficiency scores may be explained by environmental and managerial variables. Such variables have been tried in a two stage approach. A new contribution is the demonstration of how one can use the standard errors from the bias correction in stage one to improve the power of the regression model in stage two.

Five possible explanations were examined for empirical relevance, and four of them were found to be statistically significant in a multivariate weighted regression setting. More detailed data would be necessary before strong conclusions can be made, but there are indications that the most efficient building firms are characterized by high average wages, low numbers of apprentices, diversified product mixes and high numbers of hours worked per employee.

One possible problem when it comes to interpreting these results is the one of unbalanced selection. It might be that the firms that were removed from the dataset belong to a different population when it comes to the inefficiency distribution. There might be a positive correlation between entering correct data and the true technical efficiency of the units included. If the units included in the dataset are on average more efficient than the average in the population, then the overall picture of the efficiency of the industry is too optimistic.

Concerning further research a possible extension is to study time series by including data for other years. The Malmquist index could be used to decompose the productivity development of each firm into frontier shift and catching up. The relationship between productivity change and entry / exit analysis could provide additional insights.

In the current paper a bootstrapped model specification test is used to select the scale specification, but a similar approach can also be used to help select which of the inputs and outputs should be included.

It could be rewarding to examine how the weighted regression method suggested in this paper performs compared to bootstrapping in both stage one and stage two. This comparison could be done in a Monte Carlo setting.

If data on the project level became available, it could be investigated whether the findings in this paper have empirical relevance on project level data.

It would also be interesting to further investigate the theoretical relationship between the estimated bias and the original uncorrected DEA efficiency score, as well as the relationship between the estimated bias and the standard error of the bias corrected efficiency score.

References

Albriktsen, R, 1989, Produktivitet i byggebransjen i Norden, NBI Project report no. 40, Norwegian Building Research Institute.

Albriktsen, R. and Førsund, F, 1990, A productivity study of the Norwegian building industry, *Journal of Productivity Analyses*, 2-1990, pp. 53-66.

Andersen, P. & Petersen, N.C., 1993, A Procedure for Ranking Efficient Units in Data Envelopment Analysis. *Management Science*, 39(10), 1261-1264.

Banker, R.D.,1993, Maximum Likelihood, consistency and data envelopment analysis: a statistical foundation, *Management Science*, 39, 10, 1265-1273.

Banker, R.D., 1996, Hypothesis Tests Using Data Envelopment Analysis, *Journal of Productivity Analysis*, 7, 139-159.

Carroll, R.J and D. Ruppert, 1988, Transformation and Weighting in Regression, Chapman and Hall, New York.

Charnes, A., Cooper, W.W. and Rhodes, E., 1978, Measuring the efficiency of decision making units, *European Journal of Operations Research* 2, 429-444.

Cooper, W.W., L.M. Seiford, and K. Tone, 2000, Data Envelopment Analysis: A comprehensive text with models, applications, references and DEA-solver software, Boston/Dordrecht/London: Kluwer Academic Publishers.

Edvardsen, D.F., Førsund, F.R., Kittelsen, S.A.C., 2003, Far out or alone in the crowd: Classification of self-evaluators in DEA, Working paper 2003:7 from the Health Economics research program, University of Oslo.

Efron, B., 1979, Bootstrap methods: another look at the jackknife, *Annals of statistics* 7, 1-6.

Førsund, F. R. and L. Hjalmarsson, 1979, Generalized Farrell measures of efficiency: an application to milk processing in Swedish dairy plants, *Economic Journal* 89, 294-315.

Farrell, M.J.,1957, The measurement of productive efficiency, *J.R. Statis. Soc*. Series A 120, 253-281.

Groak, S. 1994, Is construction an industry? , *Construction management and economics*, 12, 4, pp 187-193.

Jonsson, J., 1996, Construction site productivity measurement: selection, application and evaluation of methods and measures. Doctoral thesis, Lulea University of Technology.

Kittelsen, S.A.C., 1993, Stepwise DEA; Choosing Variables for Measuring Technical Efficiency in Norwegian Electricity Distribution, Memo 06/1993 Department of Economics, University of Oslo

Kneip, A., Simar, L. and Wilson, P., 2003, Asymptotics for DEA Estimators in Non-parametric Frontier Models, Discussion Paper 317, Institute de Statistique, Université Catholique de Louvain.

Marron, J. S. and Nolan, D., 1988, *Canonical kernels for density estimation*, Statistics & Probability Letters 7(3): 195-199.

Ofori, G. 1994, Establishing construction economics as an academic discipline, *Construction Management and Economics*, pp 295-306, 14, 4,

Salter, W.E.G., 1960, Productivity and Technical Change, Cambridge, UK: Cambridge University Press.

Silverman, B.W., 1986, Density Estimation for Statistics and Data Analysis, published by Chapman and Hall.

Silverman, B.W. and Young, G.A.,1987, The bootstrap: to smooth or not to smooth? *Biometrika* 74, 469-479.

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., 2000, A general methodology for bootstrapping in nonparametric frontier models, *Journal of Applied Statistics* 27, 779--802.

Simar, L. and Wilson, P., 2001, Testing restrictions in nonparametric efficiency models, *Communications in Statistics*, 30, 159-184.

Simar, L. and Wilson, P., 2002, Nonparametric Tests of Returns to Scale, *European Journal of Operational Research*, 139, 115-132

Simar, L. and P. Wilson, 2003, Estimation and Inference in Two-Stage, Semi-Parametric Models of Production Processes, Discussion Paper 307, Institute de Statistique, Université Catholique de Louvain.

Torgersen, A.M., Førsund, F.R., Kittelsen, S.A.C., 1996, Slack adjusted efficiency measures and ranking of efficient units, *Journal of Productivity Analysis*, 7, 379-398.