MEMORANDUM

No 29/2005

Efficiency and Productivity of Norwegian tax Offices

Finn R. Førsund, Sverre A.C. Kittelsen and Frode Lindseth

ISSN: 0801-1117

Department of Economics
University of Oslo

This series is published by the
University of Oslo
Department of Economics

P. O.Box 1095 Blindern
N-0317 OSLO Norway
Telephone: + 47 22855127
Fax: + 47 22855035
Internet: http://www.oekonomi.uio.no/
e-mail: econdep@econ.uio.no

In co-operation with
The Frisch Centre for Economic Research

Gaustadalleén 21
N-0371 OSLO Norway
Telephone: +47 22 95 88 20
Fax: +47 22 95 88 25
Internet: http://www.frisch.uio.no/
e-mail: frisch@frisch.uio.no

List of the last 10 Memoranda:

No 28	Erling Barth and Tone Ognedal Unreported labour. 34 pp.
No 27	Marina Della Giusta, Maria Laura Di Tommaso and Steinar Strøm Who's watching? The market for prostitution services. 31 pp.
No 26	Hilde C. Bjørnland Monetary Policy and the Illusionary Exchange Rate Puzzle. 29 pp.
No 25	Geir B. Asheim Welfare comparisons between societies with different population sizes and environmental characteristics. 16 pp.
No 24	Geir B. Asheim Can NNP be used for welfare comparisons?. 24 pp.
No 23	Geir B. Asheim, Wolfgang Buchholz, John M. Hartwick, Tapan Mitra and Cees Withagen Constant savings rates and quasi-arithmetic population growth under exhaustible resource constraints. 28 pp.
No 22	Ragnar Nymoen Evaluating a Central Bank's Recent Forecast Failure. 24 pp.
No 21	Tyra Ekhaugen Extracting the causal component from the intergenerational correlation in unemployment. 22 pp.
No 20	Knut Røed and Elisabeth Fevang Organisational Change, Absenteeism and Welfare Dependency. 41 pp.
No 19	Simen Gaure, Knut Røed and Tao Zhang Time and Causality: A Monte Carlo Assessment of the Timing-of-Events Approach. 58 pp.

A complete list of this memo-series is available in a PDF® format at:
http://www.oekonomi.uio.no/memo/

EFFICIENCY AND PRODUCTIVITY OF NORWEGIAN TAX OFFICES*

by

Finn R. Førsund,
Department of Economics, University of Oslo
and the Frisch Centre
Sverre A.C. Kittelsen,
The Frisch Centre
Frode Lindseth
The Norwegian Directorate of Taxes

Abstract: The performance of local tax offices of Norway is studied over a three year period applying Data Envelopment Efficiency analysis and a Malmquist productivity index. The estimates are bias-corrected using a bootstrap approach recently developed for DEA models. The results show that bias correction and the construction of confidence intervals give a quite different picture than without bootstrapping. A set of best practice offices is identified for future work on finding explanations for good performance. The productivity development of individual offices is classified into the four categories efficient cost increase, efficient cost savings, inefficient cost savings and inefficient cost increase.

Key words: Tax office, productivity, efficiency, scale efficiency, DEA, bootstrap

JEL classification: C60, D24, L89

* The report is the result of the project "Productivity studies in the Norwegian Tax Administration" done by the Frisch Centre for the Directorate of Taxes. We are indebted to Tone Ognedal for helpful comments on a draft version and to Dag Fjeld Edvardsen for technical support and being a discussion partner during the research process.

1. Introduction

There has been an increasing emphasis in many countries in recent years on improving the functioning of the public sector. A natural first step is to measure efficiency and productivity. For the parts of the public sector not participating in markets for their products it may be a difficult task in itself just to establish the definition and measurement of outputs. Current book-keeping practices are often more designed to deal with inputs, i.e. the budget, than provide detailed enough information on the output side. For activities that produce multiple outputs with important elements of quality it can be a complex task to delimitate types of outputs and provide their measurement.

Tax offices are within the public sector in most countries. In Norway local tax assessment offices sort under the Directorate of Taxes that again sort under the Ministry of Finance. The Ministry made a direct request in 2003 to the Directorate of Taxes to work out one or more indicators of productivity in the sector. The response to this request resulted in a pilot study. The first stage was to find indicators for outputs and resources among the statistics produced by tax offices that are available for the Directorate. This process involved a number of people at the Directorate and resulted in a sharper focus on what kind of services that are actually produced and an awareness of weaknesses of the statistics and problems of comparability. A benchmark tool, data envelopment analysis (DEA), was used to calculate efficiency scores for 98 tax offices for 2002 and 2003. The main output was a ranking of offices and identification of efficient units that might serve as role models for inefficient ones. Studying the practices of such efficient units would be the starting point for finding ways of improving the resource utilisation of offices.

The present paper is a continuation of this work. The Directorate of Taxes has been especially concerned about improving the data quality and reducing any “noise” that may be creating uncertainty about the results. Since calculation of efficiency scores for individual tax offices are intended to be used to find explanations for efficiency and productivity differences, and to help identify possible measures leading to productivity improvements, it is important that the calculations are based on best available methods. A recent theoretical development of DEA is to take explicitly into account the statistical properties of efficiency scores as estimators of

unknown true scores by applying the technique of bootstrapping (see Simar and Wilson, 2000). This provides bias correction of the scores and confidence intervals, thus signalling the quality of the estimates, and especially avoiding drawing wrong conclusion as to which units should be used as role models for improvement when the density of observation is disregarded. In the pilot study this method was not available, but the methods are implemented now in the DEA software package at the Frisch Centre.

Data for one more year has been collected enabling us to also investigate productivity development using the Malmquist productivity index. The statistical technique of bootstrapping is also applied to these index values for individual offices. However, having only three years means that the results of this exercise is more useful for showing how to perform a productivity study than to provide authoritative information on productivity development.

The paper is organised in the following way. Section 2 presents the methods used with emphasis on bootstrapping. In Section 3 the data set is presented and the specification of the output and input variables that could be established. The empirical results for bias-corrected efficiency scores and productivity developments are discussed in Section 4. Section 5 concludes.

2. Methodology

Data envelopment analysis

The DEA approach is especially suitable as a benchmarking tool in a setting of multiple inputs and outputs, where price information on outputs are not available, and there is no firm knowledge about the shape of the transformation function relating outputs, y, to inputs, x. The production structure is based on the production possibility set

$$S = \{(x, y) : x \, can \, produce \, y\} \tag{1}$$

Following Farrell (1957) the production possibility set is defined empirically by enveloping the observations as tightly as possible by a piecewise linear convex outer boundary (see Banker et al. (1984) for the properties of the empirically defined set)

$$\hat{S} = \{(x,y): \sum_{j=1}^{J} \lambda_j x_{nj} \le x_n \quad (n=1,..,N) \ , \sum_{j=1}^{J} \lambda_j y_{mj} \ge y_m \quad (m=1,..,M) \ , \\ \sum_{j=1}^{J} \lambda_j = 1 \ , \ \lambda_j \ge 0 \quad (j=1,..,J)\} \tag{2}$$

where *J* is the number of observed units, *M* is the number of outputs and *N* is the number of inputs. The restriction that sum of weights, λ_j, should equal one, implies that a variable returns to scale (VRS) frontier function is specified. By dropping this requirement we will have a constant return to scale (CRS) frontier. Efficiency is measured for an observation relative to the boundary of the set S or its estimate $\hat{S}$. The boundary is termed the *production frontier*.

The boundary of the production possibility set *S* corresponds to the neoclassical notion of an efficient transformation function between inputs and outputs and will in our DEA context be termed the *frontier transformation function*. Efficiency measures for observations are in general based on the distance between an observation and the boundary. Following Farrell (1957) there are two basic directions to go from an observation to the frontier: keeping outputs fixed and moving to the frontier by a proportional reductions in inputs, or keeping inputs fixed and moving to the frontier by a proportional expansion of outputs. Corresponding to the two directions we have for a unit (*i*) the input oriented efficiency measure

$$E_{1i} = \text{Min}\{\theta : (\theta \mathbf{x}_i, \mathbf{y}_i) \in S\}, \tag{3}$$

and the output oriented measure

$$E_{2i} = \text{Min}\{1/\phi : (\mathbf{x}_i, \phi \mathbf{y}_i) \in S\} \tag{4}$$

Both measures are restricted to be between zero and one. The DEA estimate of these measures are calculated by inserting $\hat{S}$ for S:

$$\hat{E}_{1i} = Min\ \theta_i$$
$$subject\ to$$
$$\sum_{j=1}^{J} \lambda_j y_{mj} - y_{mi} \geq 0\ ,\ m = 1,..,M$$
$$\theta_i x_{ni} - \sum_{j=1}^{J} \lambda_j x_{nj} \geq 0\ ,\ n = 1,..,N \qquad (5)$$
$$\sum_{j=1}^{J} \lambda_j = 1$$
$$\lambda_j \geq 0\ ,\ j = 1$$

and

$$1/\hat{E}_{2i} = Max\,\phi_i$$
$$subject\ to$$
$$\phi_i y_{mi} - \sum_{j=1}^{J} \lambda_j y_{mj} \leq 0\ ,m = 1,..,M$$
$$\sum_{j=1}^{J} \lambda_j x_{nj} - x_{ni} \leq 0\ ,n = 1,..,N \qquad (6)$$
$$\sum_{j=1}^{J} \lambda_j = 1$$
$$\lambda_j \geq 0\ \ ,j = 1,..,J$$

The point $P_{di}^{ref} = (\sum_{j=1}^{J} \lambda_{ij} x_j, \sum_{j=1}^{J} \lambda_{ij} y_j)\ (d = 1,2)$ is on the frontier and is termed the *reference point* for unit *i*. The DEA frontier and observation and reference points are illustrated in Figure 1. The efficiency scores may be given a relative productivity interpretation. Productivity is commonly defined as the ratio between a weighted sum of outputs and a weighted sum of inputs. Assume we have weights, v_{in} (n =1,.., N) for each input and weights u_{im} (m =1,.., M) for each output for a production unit (i). The Farrell efficiency scores can then be defined as the ratio of observed productivity, P_i, and the productivity of the reference point on the benchmark frontier, P_{1i}^{ref} or P_{2i}^{ref} depending on the orientation, using the common definition of productivity as ratio of weighted outputs on weighted inputs:

$$E_{1i} = \frac{P_i}{P_{1i}^{ref}} = \frac{\sum_{m=1}^{M} u_{im} y_{im} / \sum_{s=1}^{S} v_{is} x_{is}}{\sum_{m=1}^{M} u_{im} y_{im} / \sum_{s=1}^{S} v_{is} x_{is}^{ref}},\quad E_{2i} = \frac{P_i}{P_{2i}^{ref}} = \frac{\sum_{m=1}^{M} u_{im} y_{im} / \sum_{s=1}^{S} v_{is} x_{is}}{\sum_{m=1}^{M} u_{im} y_{im}^{ref} / \sum_{s=1}^{S} v_{is} x_{is}} \qquad (7)$$

The weights in the equations above are assumed to be given numbers. When calculating the DEA models (5) or (6) the weights are actually part of the DEA output, i.e. for each unit the two sets of weights are estimated *endogenously*. (The weights are technically the shadow

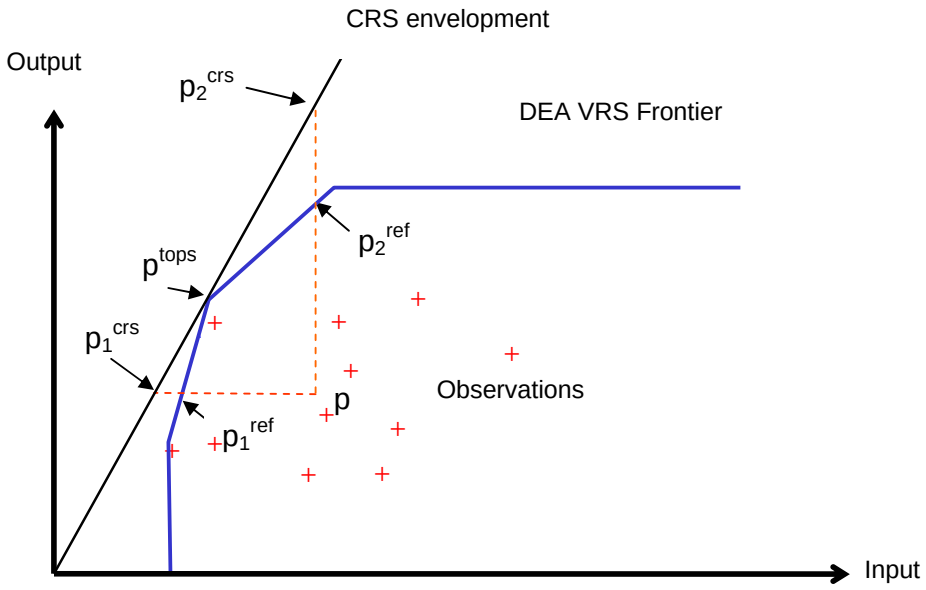

Figure 1. Definitions of efficiency measures

prices on the input and output constraints of problems (5) and (6).)

In the case of VRS a need for characterizing the observed scale of the operation arises. Three additional efficiency measures concerning the scale of the operation may be defined (Førsund and Hjalmarsson (1974), 1979). In order to calibrate scale efficiency measures the concept of technically optimal scale has to be introduced (Frisch, 1965). This is the scale where the returns to scale is one, and is illustrated in Figure 1 as the tangent point p_i^{tops} of the CRS line and the VRS frontier.[1] A measure of *technical productivity* for unit *i* is then defined as the following ratios between productivities:

$$E_{3i} = \frac{P_i}{P_i^{tops}} = \frac{P_i}{P_{1i}^{crs}} = \frac{P_i}{P_{2i}^{crs}} \tag{8}$$

The estimation of this measure is done by solving either problem (5) or (6), as indicated by the last two expressions in (8), with only nonnegativity restrictions on the weights λ. The weight used for defining productivities in (8) is obtained from the dual solution to the LP program.

[1] In general the technically optimal scale point may not be unique, i.e. the CRS line may coincide with a segment on the frontier, but the scale elasticity will be one along such a segment, see Førsund and Hjalmarsson (2004).

Measures of pure scale efficiency (scale efficiency for short) may be obtained projecting the observation radially to the frontier either in the input- or output direction and then comparing productivity with the productivity at the technically optimal scale point:

$$E_{4i} = \frac{P_i}{P_{1i}^{crs}} = \frac{E_{3i}}{E_{1i}},\ E_{5i} = \frac{P_i}{P_{2i}^{crs}} = \frac{E_{3i}}{E_{2i}},\ i = 1,..,J \tag{9}$$

The scale efficiency measures are estimated by using the corresponding estimates of technical efficiencies and technical productivities in the last equalities in each definition.

Productivity measurement

The Malmquist productivity index (Caves et al., 1982) is defined by using the efficiency scores for two different periods for a unit (or comparing two different units from the same time period) measured against the same frontier technology:

$$M_{di}^{s}(x_{iu}, y_{iu}, x_{iv}, y_{iv}) = \frac{E_{di}^{s}(x_{iv}, y_{iv})}{E_{di}^{s}(x_{iu}, y_{iu})},\ d = 1,2,\ i = 1,..,J \tag{10}$$

Here the index for the frontier technology is *s,* the index for orientation of the productivity and efficiency measures is *d*, the index for the unit is *i* and the index for the two periods is *u* and *v*. In order to calculate the efficiency scores the programme (5) or (6) must be extended to include time periods. This is straightforward: the observations used to support the frontier indexed *s* must be specified (for example observations for a specific year), and then the unit *i* from two periods is used as the observation in two separate efficiency calculations, one for each period *u* and *v*. It is then possible that the efficiency scores become greater than one.

Bootstrapping

It is well known since Farrell (1957) that a piecewise linear envelopment of data as tight as possible "from above", obeying some basic properties of production possibility sets, results in a frontier estimator that is pessimistically biased. We have a limited number of observations or realisations of an unknown technology. The situation is illustrated in Fig. 2. The input- and output orientation Farrell suggested for his technical efficiency measures are indicated by the dotted lines from the inefficient observation P. The efficiency scores are correspondingly optimistically biased. Since the DEA method is based on enveloping the observations as tightly as possible there may be potential realizations of the unknown technology that are not appearing as actual observations. The sampling bias for a given DMU can be expected to be higher the lower the number of other observations in the sample. Banker (1993) proved that as the number of draws goes towards infinity, the distance between the DEA estimate and the

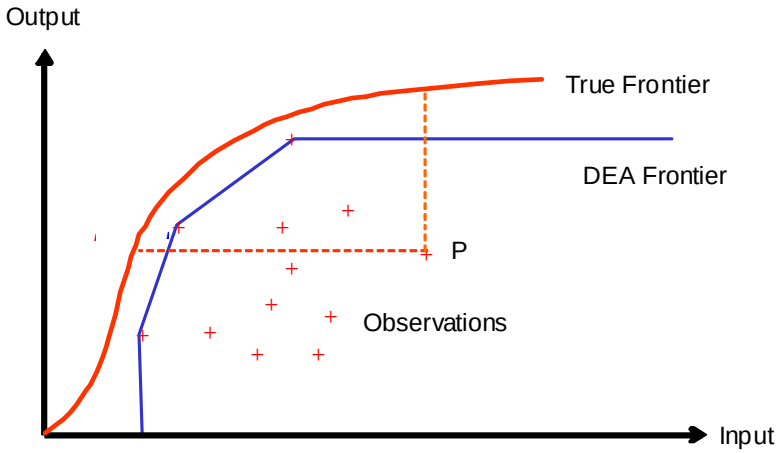

Figure 2. The inherent pessimistic bias of the DEA estimator

true efficiency score goes towards zero, i.e. the DEA estimator is consistent. 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. We know, however, that the bias is non-negative, in the sense that the DEA estimated efficiency is higher than or equal to the true efficiency.The following Data Generating Process (DGP) is assumed: observations in the production possibility set are generated by randomly drawn efficiencies from the true efficiency distribution, with exogenously given input levels and output mixes. 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. In the following a homogenous efficiency distribution is assumed, but this can be relaxed with a more complicated DEA bootstrap methodology (Simar and Wilson, 2000). 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 pseudo replicate datasets using resampling. Bias correction in DEA using bootstrapping is based on the following assumption:

$$(E - \hat{E}) \sim (\hat{E} - \tilde{E}), \tag{11}$$

where E is the true unknown efficiency, $\hat{E}$ is the original DEA estimate (see Fig. 2), and $\tilde{E}$ is the bootstrapped estimate. This estimate is obtained in the following way: The empirical distribution of the efficiency scores from the original DEA run is used to estimate a smoother

distribution by a kernel density estimate (KDE) using reflection to avoid the accumulation of efficiency score values of one (Silverman, 1986). The pseudo observation are then created by projecting all inefficient observations to the DEA frontier and drawing randomly an efficiency score for each unit from the KDE distribution. A new DEA frontier is then estimated on these pseudo observations, each generated by mimicking the original DGP, as if the original DEA estimated frontier were the true frontier. This process is illustrated in Figure 3 with a bias for one unit for iteration *i* indicated by B_i. The new frontier must lie on the inside of the original DEA frontier. We make 2000 such draws and establish 2000 new DEA frontiers, resulting in 2000 pseudo sample efficiency estimates for each observation. Using the bias correction procedure, the estimate of one point on the true frontier for each observation for each iteration No. *i* is obtained by shifting the pseudo-frontier to the left with the length of two times the bias, B_i. Once we have a value for $\tilde{E}$ we can estimate bias as $(\hat{E} - \tilde{E})$ and a bias corrected estimate of the true statistic from (9) as

$$\hat{\hat{E}} = \hat{E} + (\hat{E} - \tilde{E}) = 2\hat{E} - \tilde{E} \qquad (12)$$

This is illustrated in Figure 4. The point on the estimate of the true frontier is obtained moving the distance of twice the bias estimate to the left. The estimate of the points on the true frontier are obtained as average values from the DEA frontiers based on the bias-

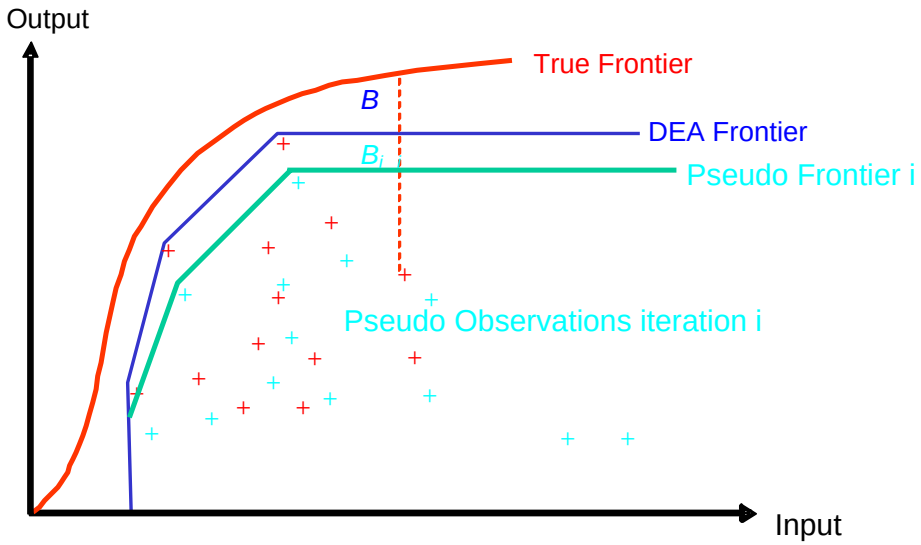

Figure 3. The bootstrap correction

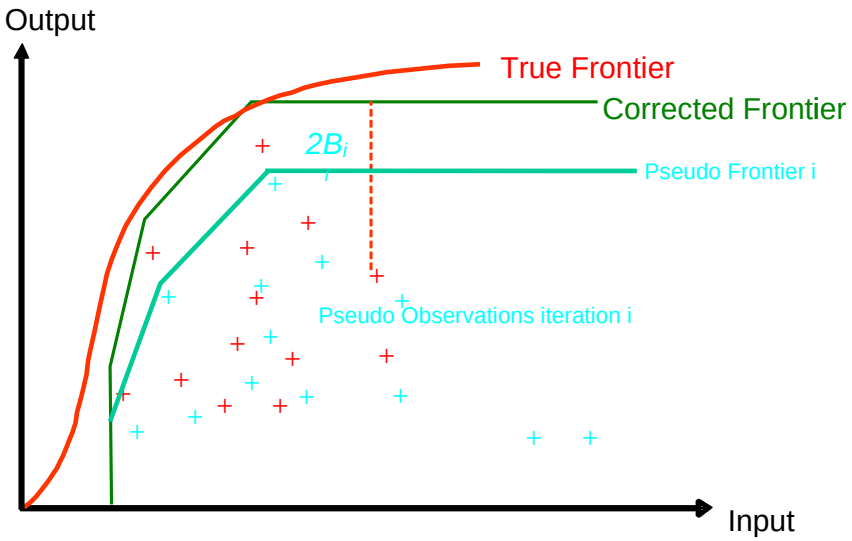

Figure 4. The Bootstrap idea
Bias correction of the efficiency score for each iteration no. i

corrected pseudo data fro the 2000 iterations. The distributions also give us the confidence intervals for the efficiency scores. The estimate of the frontier may lead to both underestimation and overestimation of the true frontier. Confidence bands are indicated by the broken curves in Figure 5.

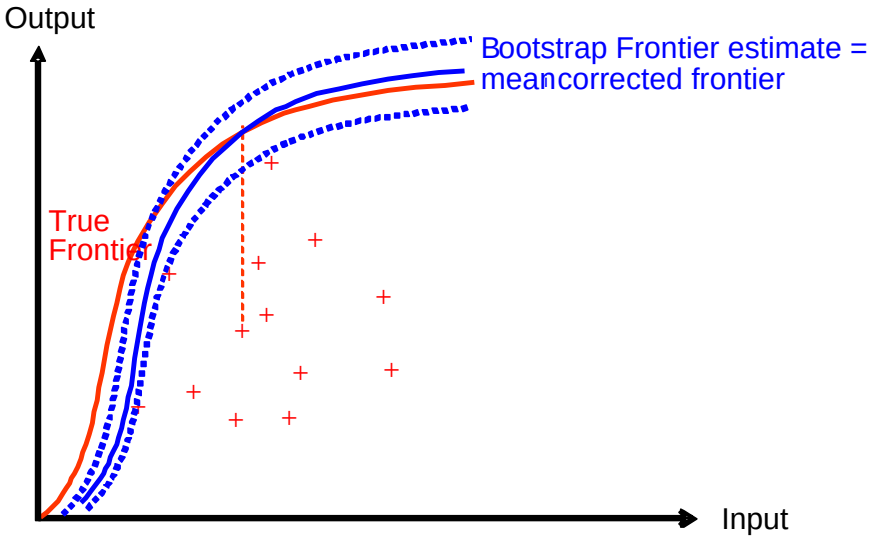

Figure 5. The Bootstrap estimate and confidence interval

3. Data

Table 1. The data

Variable	Year	Minimum	Maximum	Range	Sum	Mean	Std. Deviation
X: The cost of deployment of resources including manpower, offices and current expenses. The cost has been adjusted for compensation in the budget for special circumstances, like rent and travel costs.	2002	2 804 888	171 593 294	168 788 406	1 226 341 592	12 513 690	18 291 279
	2003	2 998 829	177 198 456	174 199 627	1 250 478 743	12 759 987	18 985 940
	2004	2 884 602	172 265 680	169 381 078	1 247 754 940	12 732 193	18 437 839
Y1: Number of people relocated during the year registered by home address and number of immigrations and emigrations.	2002	633	97 028	96 365	602 963	6 153	10 686
	2003	633	101 186	100 523	611 812	6 243	11 127
	2004	804	110 497	109 693	643 080	6 562	11 973
Y2: Number of false registrations detected by control activities.	2002	0	799	799	3 783	39	97
	2003	0	1 526	1 526	4 701	48	156
	2004	0	3 299	3 299	6 925	70	337
Y3: Number of tax returns from employees and pensioners	2002	5 361	418 785	413 424	3 384 913	34 540	46 818
	2003	5 604	422 115	416 511	3 452 177	35 226	47 531
	2004	5 601	428 822	423 221	3 462 748	35 334	48 015
Y4: Number of complaints on tax assessment	2002	40	16 295	16 255	63 407	647	1 839
	2003	9	10 018	10 009	52 573	537	1 211
	2004	9	11 178	11 169	48 680	497	1 245
Y5: Number of returns from non-incorporated businesses.	2002	801	32 510	31 709	316 542	3 230	3 411
	2003	824	33 695	32 871	325 165	3 318	3 522
	2004	791	34 722	33 931	323 610	3 302	3 669
Y6: Number of corporate tax returns.	2002	226	33 264	33 038	159 189	1 624	3 484
	2003	231	31 253	31 022	159 908	1 632	3 304
	2004	267	31 461	31 194	162 164	1 655	3 338

This empirical benchmarking exercise is restricted by the need to use pre-existing data. As mentioned in the introduction the Directorate of Taxes has had en extensive discussion about the most relevant measures for outputs and inputs. Furthermore, the data set has been controlled in several ways, e.g. finding extreme values, inspecting the distributions of variables etc., and this internal process of data control has ensured an acceptable quality for the data. The list of the variables together with some key information is given in Table 1.[2]

[2] The numbering of units has been done randomly in order to secure anonymity.

Only one input is specified; the total use of resources measured in money. The dominating expense is labour (about 80%), and the judgment in the Directorate of Taxes was that it would require too much effort to make a more detailed breakdown (e.g. labour, office, equipment, materials) comparable between units. One requirement for applying the DEA method fruitfully is that the units are using resources on the same set of outputs. In order to take into consideration that some input activities are not contributing to any of the measured outputs, the input data has been corrected by deducting items like office rent and travel expenses. This procedure is supported by the fact that the objective of the Directorate of Taxes is to find explanations for efficiency and productivity differences that in the short run can be used to improve the performances of offices. In the longer run the importance of external conditions should be studied more closely.

Six outputs are specified representing the main activity areas. Obvious output variables are the number of tax returns from individuals and returns from the two types of businesses that are specified; self-employed and limited companies. In addition one variable covering treatment of complaints and two variables covering activities checking the information about addresses are included.

We have chosen to pool the data for the 98 units for the three years for which we have observations. The input variable is adjusted using the consumer price index.

4. Empirical results

Specification test

In Section 2 we pointed out two types of technologies to use within DEA; VRS and CRS. The bootstrap procedure outlined above provides a test of which specification performs best in a statistical sense. The test is based on calculating scale efficiency. The input- and output oriented scale efficiencies E_4 and E_5 are defined in (9). Simar and Wilson (1998) suggest several test of scale specification using a bootstrapped test. They recommend the mean of the ratios:

$$\tilde{S}_2^{CRS} = \frac{1}{J} \frac{\sum_{j=1}^{J} \tilde{E}_{3j}(x_j, y_j)}{\sum_{j=1}^{J} \tilde{E}_{2j}(x_j, y_j)} \tag{13}$$

The question is whether average scale efficiency we calculated using uncorrected DEA 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.g. E_5. If the average E_5 we originally calculated using DEA is outside the given density range, e.g. 95%, we then choose to discard the H_0 that "The true technology exhibits CRS" and use VRS instead. The result of the test is shown in Figure 6. The estimated value is close to 0.94 as indicated by the broken line to the left, and thus much less than the critical value of the test. CRS is therefore rejected and VRS adopted.

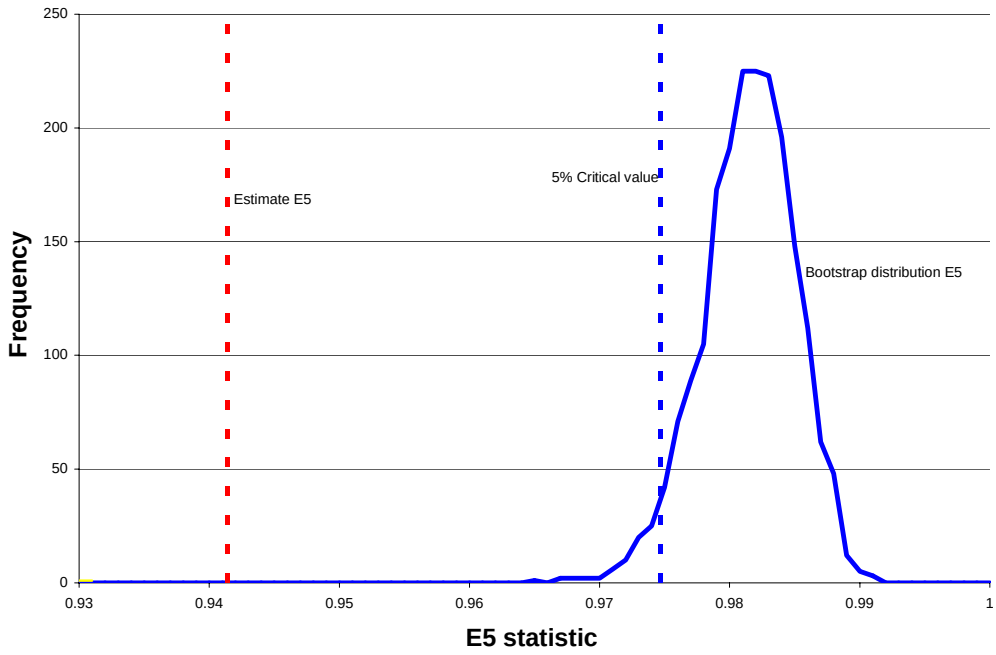

Figure 6. Specification test

The bias-corrected efficiency scores

As described in Section 2 the first step of the bootstrapping procedure is to estimate the KDE efficiency score distribution. The original distribution and the smoothed KDE distribution are shown in Figure 7a-b. The uncorrected distribution has a high number of units (38) with a score of 1, while the KDE distribution has no unit as fully efficient.

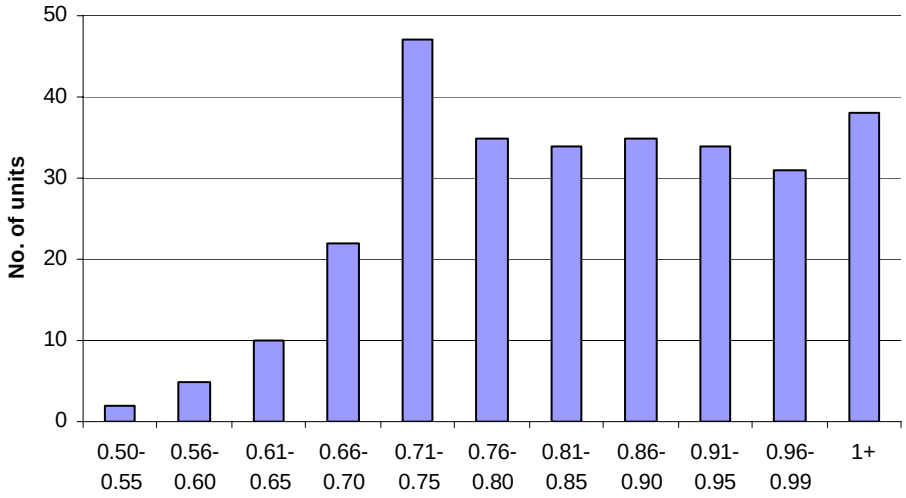

Figure 7a. The empirical frequency distribution for the initial DEA run. Output oriented efficiency score

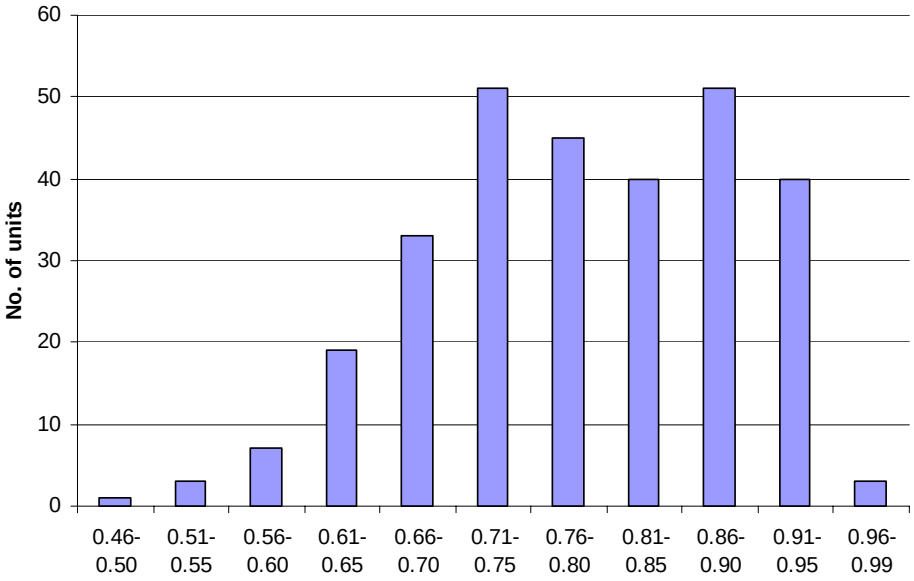

Figure 7b. The smoothed frequency distribution for the initial DEA run applying KDE with reflection. Output oriented efficiency score

But the distribution has kept its bimodal shape. Using this distribution to draw efficiency scores generate the bootstrap estimates, as explained in Section 2.

The bias corrected scores are shown in Fig. 8 together with the uncorrected scores (obtained from the first step DEA run shown in Figure 2). Remember that each unit is represented three times, once for each year. Each bar represents a unit, and the width of a bar is proportional to total costs used as the input. The units are sorted according to increasing value of the efficiency score. The left part of the distribution, representing the least efficient units (having about 1/4 of total input costs), contains almost exclusively small units. Large units tend to be localised in the middle of the distribution, while medium sized and somewhat smaller units represent the most efficient ones to the right of the distribution.

The (arithmetic) average value of the efficiency score went down from the original estimate of 0.83 to a bias corrected estimate of 0.78. This may not seem as such a dramatic change, but it is the *distribution* of change that should be observed. The bias correction is appearing as a

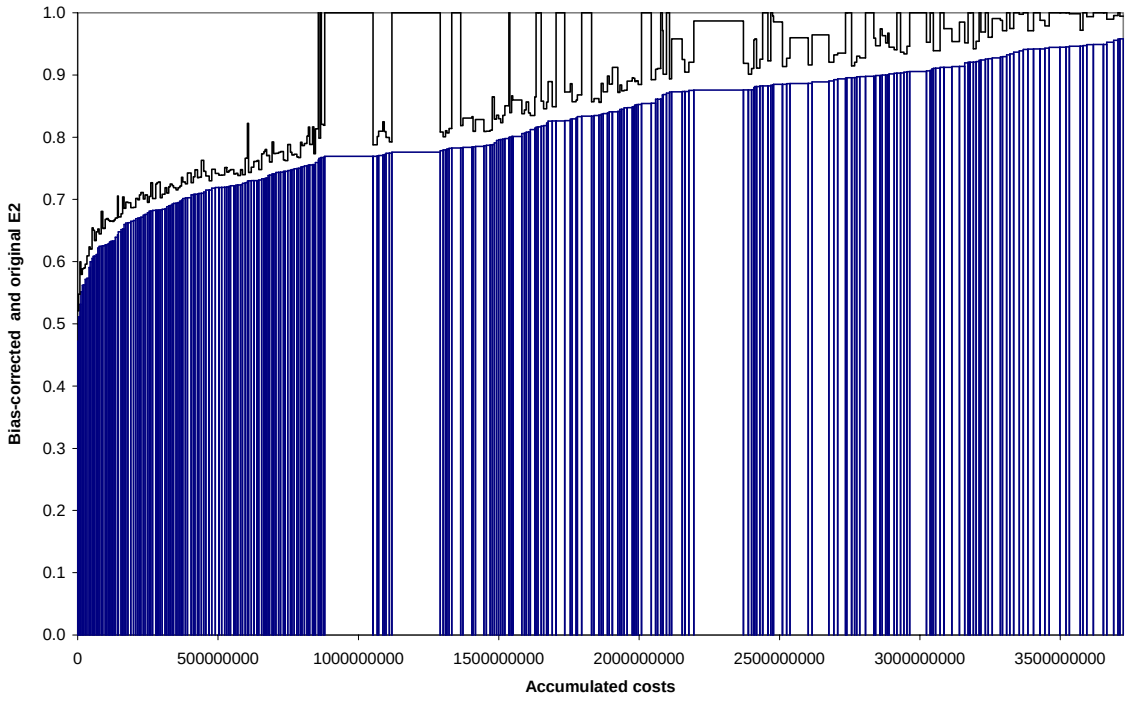

Figure 8. Sorted efficiency distribution of bias corrected output oriented scores. Uncorrected scores as a step plot.

systematic downward shift of the efficiency scores for the least efficient units representing about 1/4 of total input costs. For the larger units the difference between uncorrected and bias-corrected scores is dramatic. Many of the larger units that are fully efficient in the original DEA run get a considerable downwards bias correction. Thus bootstrapping and bias correction is very important for the selection of units to serve as role models for inefficient units.

The confidence intervals around the bias corrected estimates of the efficiency scores for all observations are shown in Figure 9. The units are sorted according to increasing average values. The impression from Figure 8 is confirmed for the least efficient tale comprising about 25% of all observations. The confidence intervals are the smallest at this end of the distribution, implying that the ranking of these units is rather reliable. It is clearly shown that most of the units having efficiency scores of 1 in the initial DEA run have the widest confidence intervals. Some of the large units are among these units. But notice that for the upper best practice tail of the distribution the confidence intervals get markedly narrower, although not so narrow as for the least efficient tale.

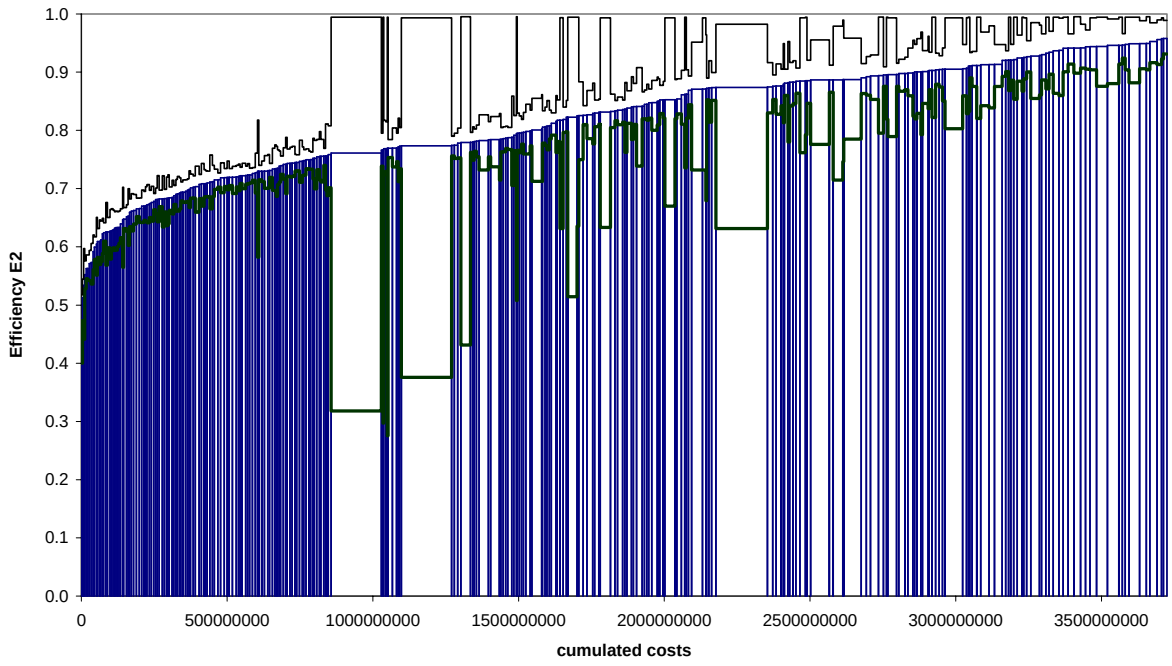

Figure 9. Biased corrected output-oriented efficiency scores
All units 95% confidence intervals

Best practice units

The initial uncorrected DEA run yields the set of efficient units (efficiency score of 1) and the set of inefficient units (efficiency score less than 1). The efficient units are termed *Peers* in DEA because such units define the reference point on the frontier for each inefficient unit. If one stops at calculating just the initial DEA model, then the peers relevant for each inefficient unit will be natural role models for the inefficient units studying how to improve their performance. One main advantage of the DEA method has, in fact, been considered to be the identification of peers. However, when the efficiency scores are bias-corrected this direct link between role models and the inefficient units disappear. It is only in the initial DEA run that the peer concept is uniquely defined. The application of the bootstrap bias correction of the efficiency scores may even most commonly result in no unit being fully efficient. But there is still the need for identifying a set of units serving as role models. We have chosen to pick out high-performing units that we will call *best practice* (BP) units. The selection criterion has been to pick the units with the highest average bias-corrected output-oriented efficiency score for the three years. A sample of 10% of the units have been selected, i.e. 10 units. The highest efficiency score is 0.96. The efficiency score for the individual years range from 0.96 to 0.88. The largest of the 10 BP units is the third largest in the full data set with about 130,000 tax returns (in 2002), while most of the other units are of about average size, ranging from 30,000 to 80,000 tax returns (in 2002). A comparison of the size structure of the BP units compared with the total average values (averaging over 3 years and 98 units) can be seen from Figure 9. While the average BP unit uses 46% more of the inputs than the sample average, it produces 75% more registration of people that have moved, 69% more tax returns from employees and pensioners and 62% more tax returns from firms. Of the three remaining outputs tax returns from personal businesses are 28% higher and number of complaints 10% higher, while the number of wrong addresses detected is 5% lower than for the average unit. So the structure of the BP units indicates that it is the relatively high number of tax returns from employees and pensioners and from firms that are the main characteristics driving the results.

The results for the 10 best practise units are illustrated in Figure 10. The best practice units are sorted according to decreasing average bias-corrected efficiency. Results for each of the three years are shown in chronological order. The span of the values of the confidence intervals is rather wide; between 0.99 and 0.85. The two first units have stable values, but the confidence intervals are wide. The values of other units vary more, although there is a

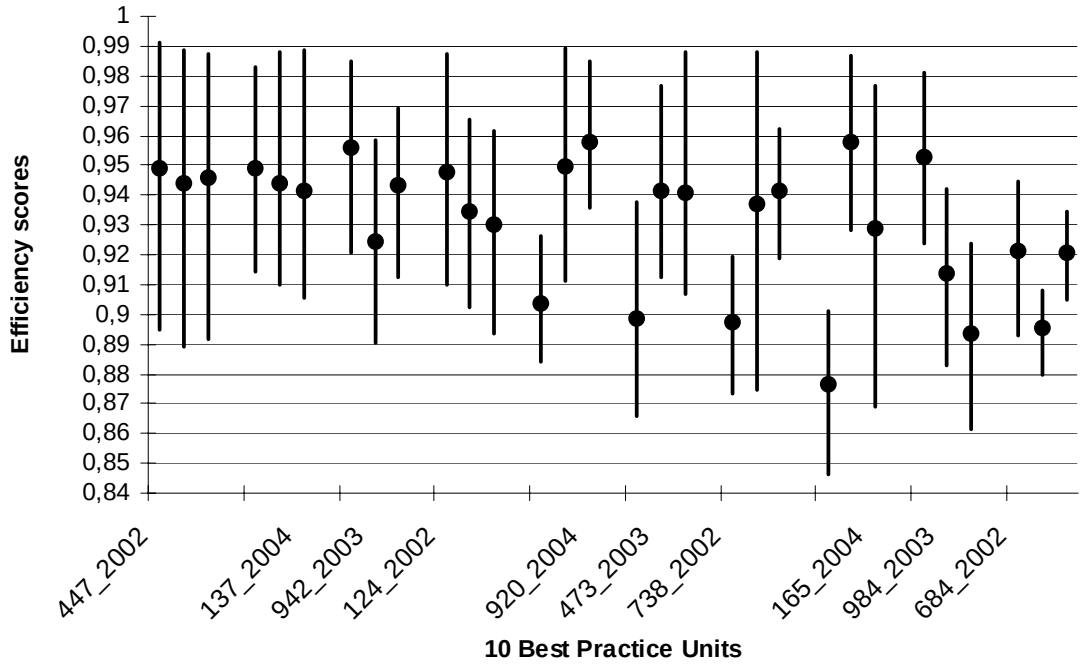

Figure 10. The 10 Best Practice units each year.
Bias corrected efficiency scores and 90% confidence interval limits

tendency for two of the three years to be more like each other. Based on the confidence intervals only a few of the observations have significantly different (better or worse) efficiency levels than the other BPs. The first unit is not significantly more efficient than any of the other BP units on the 10% confidence level. Unit 920 observed in 2004 is significantly more efficient than six other observations, while unit 165 in 2002 is significantly less efficient than 12 other observations, including itself in 2003.

It may be interesting to compare the performance of the 10 selected best practice units with their role in the initial DEA run. The peer units and their peer index values, which lie in the interval between 0 and 1 and show their importance as peers (Torgersen et al., 1996) are set out in Table 2. Of the chosen 10 best practice units three do not appear as peers, while two units (the two highest ranked ones) appear twice and the other five once. The most influential unit is no. 755 in 2003 and 2004, associated with over 20% and 14% respectively of the output increasing potential of the inefficient units they are referencing. This unit is of somewhat less than average size (40,000 tax returns in 2002); 28,000, but close to the median size of 27,000. The unit in 2003 is referencing 96 and the 2004 - unit 109 inefficient units.

Table 2. The peer index (output increasing potential referencing shares) according to the initial DEA solution.

Outputs						
Peer Units	y1+	y2+	y4+	y6+	y7+	y9+
938_2002	0.002271	0.001072	0.004348	0.007647	0.002968	0.002631
823_2002	0.00444	0.001197	0.002595	0.001728	0.002615	0.00283
212_2002	0.007206	0.008641	0.004889	0.003636	0.003208	0.006465
693_2002	0.004467	0.022551	0.001361	0.004009	0.001072	0.001878
124_2002	0.00612	0.026365	0.008017	0.014094	0.010159	0.008418
293_2002	0.064306	0.07966	0.053827	0.054425	0.042682	0.060425
265_2002	0.043647	0.030449	0.059208	0.034525	0.057622	0.049679
447_2002	0	0	0	0	0	0
628_2002	0.002003	0.000278	0.002548	0.002072	0.003003	0.00183
320_2002	0.002511	0.009337	0.003073	0.012487	0.002153	0.00389
920_2003	0.000063	0.00213	0.000083	0.000038	0.000592	0.000075
430_2003	0.00635	0.001634	0.005981	0.006891	0.005653	0.00607
911_2003	0.019031	0.026827	0.02162	0.023144	0.018741	0.023294
738_2003	0	0	0	0	0	0
750_2003	0.028117	0.012226	0.029218	0.034536	0.025114	0.019227
820_2003	0.03211	0.018344	0.038265	0.018537	0.035267	0.026124
137_2003	0.04339	0.032886	0.029557	0.02782	0.022466	0.031371
451_2003	0.018903	0.009464	0.020914	0.016298	0.020608	0.019604
447_2003	0.000971	0.010151	0.001044	0.003997	0.003731	0.006455
755_2003	0.220237	0.134995	0.209605	0.208036	0.225141	0.206686
165_2003	0.062138	0.067439	0.074042	0.07008	0.061428	0.065161
93_2003	0.000233	0.000101	0.000302	0.000199	0.00076	0.000275
572_2003	0.004812	0.0022	0.007508	0.00368	0.00802	0.006853
320_2003	0.010961	0.005326	0.004963	0.003912	0.006384	0.008994
473_2004	0.024685	0.027053	0.017876	0.039779	0.014601	0.032477
430_2004	0.061657	0.035189	0.049465	0.059335	0.040226	0.048888
34_2004	0.000376	0.00017	0.000515	0.000528	0.000496	0.000464
212_2004	0.042352	0.0362	0.036088	0.071485	0.028244	0.043455
693_2004	0.026089	0.168893	0.007708	0.031276	0.007811	0.012415
137_2004	0.009889	0.013586	0.010862	0.013366	0.00997	0.011479
77_2004	0.001534	0.003254	0.000508	0.000251	0.001477	0.001971
662_2004	0.023225	0.013818	0.011583	0.0141	0.016549	0.035375
451_2004	0.00665	0.003768	0.008794	0.003522	0.009853	0.008131
755_2004	0.132745	0.132704	0.164232	0.149232	0.165608	0.146966
93_2004	0.00673	0.014574	0.008286	0.011041	0.01198	0.01101
572_2004	0.076175	0.043778	0.096544	0.052145	0.127648	0.084247
373_2004	0.002911	0.003389	0.003751	0.001775	0.004666	0.003695
851_2004	0.000693	0.000351	0.000819	0.000374	0.001481	0.001187

This reflects the central position of this unit in the data. The third most influential unit is No. 572 in 2004 with an average peer index of 8% and referencing 85 units. But these three most influential observations are not among the sample of 10 BP units. Applying bootstrapping definitely gives another picture of which units that should serve as role models than

traditional DEA. However, the fourth most influential unit with an average peer index of 7% and a count of inefficient units of 34 is among the 10 BP units. This unit is also of median size. Of the remaining best practice units eight observations also appear as peers, but with small peer index values. Indeed, two BP observations appear as self evaluators, usually indicating some extreme values or combination of values of the input and output variables. This may be the case for unit 447 in 2002 since it is the third largest unit. The other self evaluator, unit 738 in 2003, is also larger than average size, but may be an interior self evaluator (see Edvardsen et al., 2003).

Scale efficiency

As pointed out in Section 2 the efficiency measures have a relative productivity interpretation. A reference of special interest is the maximal productivity on the frontier. From production theory we know that this maximal productivity is characterised by constant returns to scale. The bias-corrected E_3 measure set out in Figure 11 shows the observed productivity relative to the maximal productivity, keeping the same proportion between the outputs as observed, sorted in ascending order. The relative size of the units is measures by

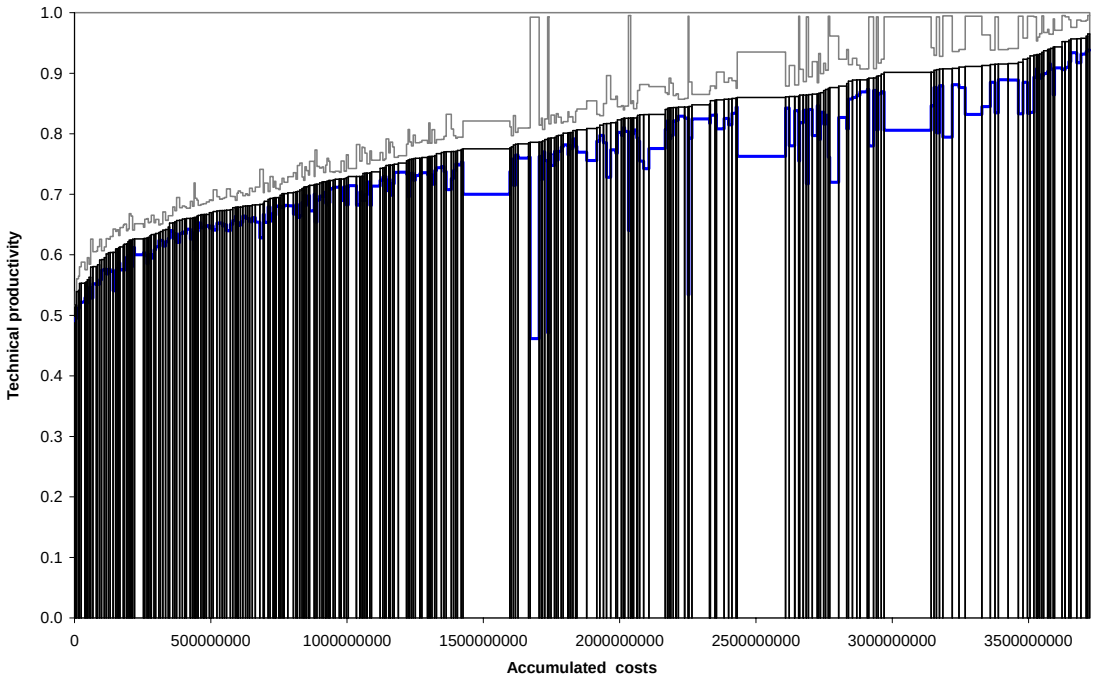

Figure 11. Efficiency relative to optimal scale (technical productivity)
95% confidence intervals

input costs (accumulated for the three years), and the upper and lower limit of the 5% confidence intervals are shown as step curves. No unit has maximal productivity, but the upper tail of the most efficient units comprising about 20% of total costs, covers the interval 0.90 to 0.97. The units are mainly medium sized and small with the exception of one large unit. It is interesting to note that 2/3 of the BP units are in this interval, and that the confidence intervals for technical productivity generally are narrower than for technical efficiency.

The worst practice tail on the interval 0.49-0.77 comprising about 38% of costs consists almost exclusively of small units (interval ending at one large unit). When we look at the confidence intervals of this worst practice tail they have a rather even and limited range compared with confidence intervals of the units with higher bias-corrected technical productivity. For some of the remaining units performing better the confidence intervals are rather wide, even so wide as to comprise the value 1. This means that a hypothesis that the unit has maximal technical productivity cannot be rejected. The implication of the findings of the technical productivities is that productivity can be improved mainly by small units becoming larger.

The technical efficiency scores pictured in Fig. 8 do not tell us whether units with lower productivity than maximal are efficient or not. Scale efficiency (pure) was defined in Section 2. This measure consists of a ratio, but the upper bound on the value is 1. In the case of output orientation, bootstrapping this ratio, E_5 (= E_3/E_2), the way we have bootstrapped the measures E_2 and E_3 leads to many of the upper limits of E_5 exceeding 1. However, the scale test used above and introduced in Simar and Wilson (1998) does not bootstrap E_5 in this way, and may be give better estimate of the distribution of E_5. We have therefore chosen to use the results for the initial run and establish confidence intervals without shifting the DEA frontiers the distance of two biases to the left. In any case the scale inefficiencies are quite limited compared with output-oriented efficiency and technical productivity scores, ranging from 0.8 to 1. Upper confidence intervals imply that for over 2/3 of the units a hypothesis that the units are scale efficient cannot be rejected. Units representing 20 % of total costs have upper confidence intervals ranging from 0.9 to 1. The implication is that if the units manage to move to the frontier using the same amounts of inputs most of the potential productivity improvements shown in Fig. 11 will be realized.

Productivity development

Due to the short time span we have data for and lack of information about development of frontier technology for tax offices we have assumed that the technology is the same for all years. This means that when we measure the productivity development for an office it is the change in *efficiency* relative to the optimal scale that will constitute the productivity change. In the definition of the Malmquist index (8) the technology index s refers to the pooled sample, and the years u and v for a unit may be bilateral combinations of the years 2002, 2003 and 2004. The productivity development for the units between 2002 and 2004 are set out in Figure 12. Since there are only three years of observation one should be careful interpreting trends. The productivity change distribution ranges from a 20% decline to a 35% increase. Taken at face value the results indicate that units representing about half of the costs (in 2004) have had a productivity decline over the three years, while a half have had a productivity improvement, the latter being somewhat more pronounced than the former. This can be visualized by comparing the "triangles" formed by the areas between the end of the histograms for productivity changes and the line of 1, i.e. the triangle on the left below the line for productivity decrease and to the right above the line for increase. The average (unweighted

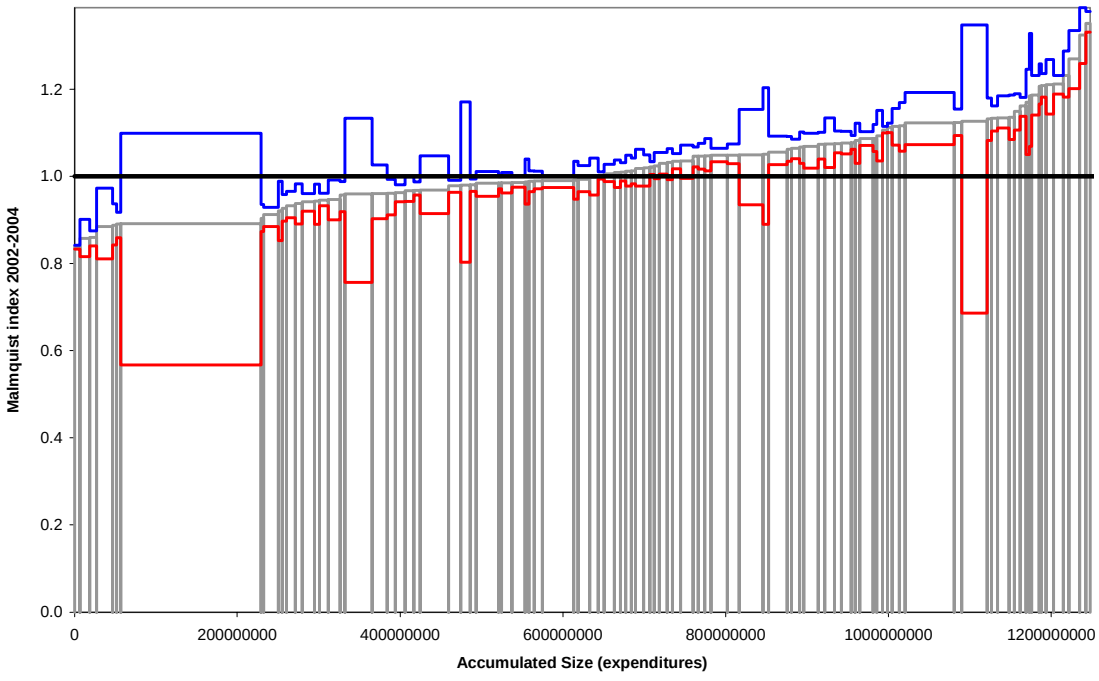

Figure 12. Malmquist productivity index 2002 – 2004.
Bias-corrected output-oriented scores

arithmetic) is a growth of 4%. Among units with productivity improvement the small ones dominate. Some average sized units have had slight improvements while others have experiences decline.

The confidence intervals shown as step curves in Figure 12 show a pattern of relatively small intervals for small sized units while the intervals for the larger units tend to be much wider with a few exceptions. The implication is that we can trust the results for the small units, but that we must be careful when using productivity figures for the large units. More reduced data density in the neighbourhood of large units make the determination of the productivity score more uncertain. Some of the medium and large units with productivity decline have upper limits of the confidence interval above 1, so a hypothesis that these units have experienced productivity growth cannot be rejected.

A further characterisation of the nature of productivity growth can be made by comparing the change in the resources used and the productivity score (see Førsund and Kalhagen, 1999). In Figure 13 a scatter diagram of productivity change from 2002 to 2004 is shown together with the relative change in input costs. To the left of the origin costs have decreased from 2002 to 2004 while to the right costs have increase. The total range is from -20% to +23%. Together

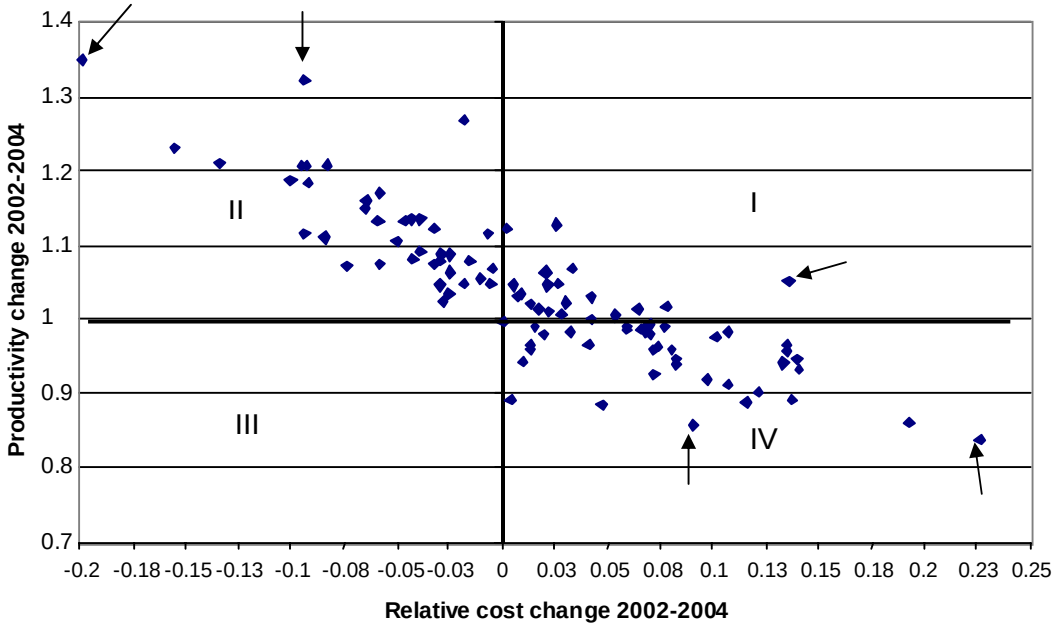

Figure 13. Productivity and cost change 2002-2004

with the horizontal line of 1 delimitating the units with productivity decrease and increase the lines form four quadrants numbered I to IV. In quadrant I units have had both productivity growth and cost increase. Such units may be said to have experienced *efficient cost increase*. The unit (indicated by an arrow) with the highest cost increase has had a productivity growth of 5% and has expanded the costs with 14%. The units in quadrant II have also had productivity growth, but experienced cost reductions. This may be termed *efficient cost savings*. The unit with the highest productivity change has had an increase of 35% (maximal of all units) and reduced the costs with the maximal cost savings of 20%. Another unit with the second highest productivity growth of 32% has had a cost decrease of 9%. In quadrant III productivity decrease is combined with cost decrease. This is *inefficient cost savings*. There are no units in this quadrant. Units in quadrant IV have the worst of both worlds with decreasing productivity and increasing costs. This is *inefficient cost increase*. The unit with the highest productivity decline, 14%, has had the maximal cost increase of 23%.

5. Policy conclusions

The main objective of performance measurement of units is to characterise performance in such a way that ways of improving performance can be found. This is of especial importance for a public service production sector not selling the services in a market. The present study has shown that is of crucial importance to use methods that enables us to make a statistical assessment of the efficiency estimates that are the “engine” of performance measurement. The importance of the data density for different dimensions of the data space is made explicit. The results show that large units may easily appear with a better performance than they should. In numerous DEA studies without bias correction efficient units are identified as role models without much qualification. Performing bias correction using bootstrap techniques enable us to estimate confidence intervals for efficiency scores and establish new criteria for selecting role models or Best Practice units. Such units can then be studied carefully in order to reveal their efficiency and productivity “secrets”.

Since small units are more numerous, the efficiency and productivity results for them are determined with greater accuracy. It is revealed that small units tend to belong to the least efficient part of the efficiency distribution. This is especially the case for technical

productivity implying that if a structure of small units is wanted, then a price to pay is reduced productivity. But the results for output-oriented efficiency and scale efficiency show that if the small units can expand their outputs to the frontier using the same inputs, then most of the scale inefficiencies vanish. Our results give a potential output increase between 21 and 24 % if all units in all periods become efficient. The additional productivity gain by becoming of optimal size is probably considerably less, but the exact effects on outputs and inputs are complex calculations because both outputs and inputs have to change. In addition maximal productivity may not only be achieved changing outputs proportionally, but also by changing output mix (Førsund and Hjalmarsson, 2004). Therefore such calculations require a separate investigation.

When interpreting the measures of productivity over time the very limited number of years (3) should be taken into consideration. Taken at face value units representing half the costs have had productivity decline and half have had productivity increase in the range of -20% to +35%, but resulting in an overall productivity increase of about 4%. The range of change may seem somewhat surprising for such a short period. For any policy actions it should be noted that the confidence intervals for the large units are very wide, while they are narrow for small units.

The type of performance evaluation performed in this study reveals inefficiency and productivity structures, but does not provide ready explanations of causes for the revealed differences. A standard procedure in the literature has been to investigate possible explanatory variables by regressing efficiency or productivity scores on candidates for explanatory variables, i.e. variables in addition to the ones used as inputs or outputs. But such a two stage procedure is not statistically satisfactory if the efficiency scores are not bias-corrected as done by the bootstrap procedure employed in this paper. One way to proceed in the second stage is to weigh the efficiency score estimates with standard deviations to make the regression more efficient (Edvardsen, 2004). Another way is to integrated bootstrapping and regressions as done in Simar and Wilson (2005).

References

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

Banker, R.D., A. Charnes and W.W. Cooper (1984) "Some models for estimating technical and scale inefficiencies." *Management Science* 30, 1078-1092.

Caves, D.W., L.R. Christensen and E. Diewert (1982): “The economic theory of index numbers and the measurement of input, output, and productivity,” *Econometrica* 50(6), 1393-1414.

Edvardsen, D. F. (2004): *Four essays on the measurement of productive efficiency*, Ph.D. thesis, Department of Economics, School of Economics and Commercial Law, Göteborg University.

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

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.

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

Frisch, R. (1965): *Theory of production,* Dordrecht: D. Reidel Publishing Company.

Førsund, F. R. and L. Hjalmarsson (1974): “On the measurement of the productive efficiency,” *Swedish Journal of Economics* 76, 141-154.

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

Førsund, F. R. and L. Hjalmarsson (2004): “Are all scales optimal in DEA? Theory and empirical evidence,” *Journal of Productivity Analysis* 21(1), 25-48.

Førsund, F. R. and K. O. Kalhagen (1999): “Efficiency and productivity of Norwegian colleges,” in G. Westermann (ed.), *Data envelopment analysis in the service sector,* Wiesbaden: Deutscher Universitäts-Verlag, 1999, 269-308.

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

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 (2000): "Statistical inference in nonparametric frontier models: the state of the art," *Journal of Productivity Analysis* 13, 49-78.

Simar, L. and P. W. Wilson (2005): "Estimation and inference in two-stage, semi-parametric models of production processes," *Journal of Econometrics*, Article in press.

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.