MEMORANDUM

No 14/2009

Productivity of Tax Offices in Norway

Finn R. Førsund
Dag Fjeld Edvardsen
Sverre A. C. Kittelsen
Frode Lindseth

ISSN: 0809-8786

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

Last 10 Memoranda

No 13/09	Johan Eyckmans and Snorre Kverndokk *Moral Concerns on Tradable Pollution Permits in International Environmental Agreements*
No 12/09	Diderik Lund *Marginal versus Average Beta of Equity under Corporate Taxation*
No 11/09	Fridrik M. Baldusson and Nils-Henrik M. von der Fehr *Price Volatility and Risk Exposure: on the Interaction of Quota and Product Markets*
No 10/09	Dag Morten Dalen, Enrico Sorisio and Steinar Strøm *Choosing among Competing Blockbusters: Does the Identity of the Third-party Payer Matter for the Prescribing Doctors?*
No 09/09	Ugo Colombino, Erik Hernæs, Marilena Locatelli and Steinar Strøm *Towards and Actuarially Fair Pension System in Norway*
No 08/09	Kjell Arne Brekke, Karen Evelyn Hauge, Jo Thori Lind and Karine Nyborg *Playing with the Good Guys: A Public Good Game with Endogenous Group Formation*
No 07/09	Benedicte Carlsen and Karine Nyborg *The Gate is Open: Primary Care Physicians as Social Security Gatekeepers*
No 06/09	Alessandro Corsi and Steinar Strøm *The Premium for Organic Wines? Estimating a Hedonic Price Equation from the Producer Side*
No 05/09	Jo Thori Lind, Karl Moene and Fredrik Willumsen *Opium for the Masses? Conflict-induced Narcotics Production in Afghanistan*
No 04/09	Jo Thori Lind and Karl Moene *Misrely Developments*

Previous issues of the memo-series are available in a PDF® format at:
http://www.oekonomi.uio.no/memo/index.html

PRODUCTIVITY OF TAX OFFICES IN NORWAY [*]

by

Finn R. Førsund

Department of Economics, University of Oslo

Dag Fjeld Edvardsen

SINTEF Building and Infrastructure, Oslo, Norway

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 using Data Envelopment Efficiency analysis and calculating Malmquist productivity indices. One input, labour, is used, and six output categories of the main service activities carried out by tax offices are specified. A bootstrap approach recently developed for DEA models is applied to establish confidence intervals for the individual indices enabling an identification of units that have either significant productivity decline or growth, or no change. A specially developed graphic display gives a visual test and grouping into the three possible categories. Looking at change in labour use and productivity change together the productivity development of individual offices is classified into the four categories efficient labour increase, efficient labour savings, inefficient labour savings and inefficient labour increase.

Key words: Tax office, Malmquist productivity index, DEA, bootstrap

JEL classification: C60, D24, L89

[*] The paper is written within a project, undertaken by the Frisch Centre for the Directorate of Taxes, on evaluating whether a recent reorganisation of the Directorate leads to a more efficient use of resources.

1. Introduction

The public Directorate of Taxes of Norway has recently undergone a comprehensive reorganisation involving creating new units for dealing with the various tasks of the Directorate, and new ways of organising the tasks. The 99 local tax assessment offices that used to sort under 20 counties, have been changed into considerably fewer offices sorting under five regional units. The tasks of these offices fall into five main categories; providing information and service to the public at large, assessing taxes, control and legal issues, tax evasion, and collecting taxes. A main feature of the reorganisation is that the offices can deal with cases originating at any locality in Norway. Tax returns from the capital Oslo can now be assessed, e.g., by an office located in Northern Norway. The idea is to exploit economies of scale and scope.

The Directorate of Taxes is interested in evaluating the effects of the reform concerning the efficiency and productivity of the new organization. One difficulty is that the basic units before and after the reorganisation change. However, these new units are still producing the same services and can be modelled using the same input-output description. Therefore, it makes sense to calculate the productivity under the old regime as one part of doing the evaluation. The objective of the present paper is to calculate the productivity change of the tax offices that existed before the reform, as well as to establish a reference for evaluating future performance changes.[1]

Productivity is normally perceived as the ratio of outputs to inputs, and in the presence of multiple inputs and outputs these must be weighed together to single numbers for outputs and inputs, respectively, in order to measure total factor productivity (TFP). There are two main strategies to follow regarding creation of weights (Førsund, 1997). The index approach utilises exogenous information; the standard weights are the prices of outputs and inputs. Well known such productivity indices are the Törnqvist index and the Fisher index. The technology approach

[1] Previous connected research is presented in Førsund et al. (2005) and further elaborations in Førsund et al. (2006a,b). New features in the present paper are that total costs, used as a single input in the previous studies, has been substituted with labour as input, leading to somewhat different results, and new ways of showing graphically the implications of confidence interval and nature of productivity change have been devised.

is based on information about the production technology, and weights are explicitly or implicitly deduced from the technology. It should be noticed that when some or all output prices are lacking, such as often is the case in public sector service production, only the technology approach has been used. Our approach is based on the technology approach of estimating the transformation of inputs into outputs. Due to lack of knowledge of functional forms and consequently the need of letting the data speak maximally concerning the nature of the transformation, a non-parametric method termed Data Envelopment Analysis (DEA) is employed.[2]

An important issue of evaluations of performance is whether changes are significant or not. A recent report from a British Working Party of Performance Monitoring in the Public Services has, as one of the recommendations that reported performance measures should always include measures of uncertainty (Bird et al., 2005). Since the intention of calculating productivities for individual tax offices is to use them as a benchmark for performance after the reorganisation of the Directorate, it is important that the calculations are based on best available methods. A recent theoretical development of the DEA method is to take explicitly into account the statistical properties of efficiency scores as estimators of unknown true scores by applying the technique of bootstrapping (Simar and Wilson, 1998, 2000). This overcomes the inherent sampling bias of limited data sets under certain assumptions. Bootstrapping provides bias correction of the scores and confidence intervals, thus signalling the quality of the estimates of productivity levels and changes.

Data has been collected especially for the study for three years enabling us to investigate productivity development using the Malmquist productivity index Caves et al., 1982). The statistical technique of bootstrapping is applied to these index values for individual offices. The productivity change distribution for the total period ranges from a 40% decline to a 45% increase. Taken at face value the results indicate that units representing about 40% of the labour force in 2004 have had a productivity decline over the three years, while 60% has had a productivity improvement. The confidence intervals tend to be wider for large units. Key

[2] The non-parametric method was originated in Farrell (1957) and further developed into the tool in use today in Charnes et al. (1978), where the name DEA was coined.

information about uncertainty is provided by testing whether change is significantly positive or negative. About 17 % of the units had a significantly declining productivity, and 47 % significantly increasing productivity. The productivity development of individual offices, based on bias-corrected productivity measures, is classified into the four categories efficient labour increase (I), efficient labour savings (II), inefficient labour savings (III) and inefficient labour increase (IV). Most units belonged to the groups II and III, while very few units belonged to group I.

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

2. Methodology

The point of departure for defining the production technology that will be the basis for measuring productivity is to formulate a production possibility set which can accommodate inefficient as well as efficient operations. Let *x* be a vector of inputs and *y* be a vector of outputs, then the production possibility set at time *t* is defined as:

$$S^t = \left\{ (x,y) \middle| \, x \text{ can produce } y \text{ at time } t \right\} \tag{1}$$

In the presence of inefficient operations the relevant technology reference is the efficient border of the technology set. This border set will be termed the *frontier production function*. Following Farrell (1957) the production structure is based on a convex production possibility set $\hat{S}^t$, as an estimator of the true production possibility set S^t in (1), 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).

Productivity measurement

Building on the idea in Malmquist (1953) of proportional variation of variables when measuring change, Caves et al. (1982) introduced the bilateral Malmquist productivity index based on the ratio of Farrell (1957) efficiency measures for the two units (e.g. the same unit measured for two different time periods). Efficiency is measured against the same frontier technology (*s*). The Malmquist productivity index (Caves et al., 1982) is developed for discrete time and defined by using the Farrell efficiency scores for two different periods (*u* and *v*) for a unit (*i*). The Caves et al. definition of the Malmquist productivity index is:

$$M_{di}^{s}(u,v)=\frac{E_{di}^{s}(x_{iv},y_{iv};S^{s})}{E_{di}^{s}(x_{iu},y_{iu};S^{s})}\ ,\quad d=1,2,\ i=1,..,J,\ \ s,u,v=1,..,T,u<v \qquad (2)$$

Here the index for the frontier technology is *s*, the index for the orientation is *d* (*d* = 1 is input orientation, *d* = 2 is output orientation), the index for the unit is *i*, the index for the two time periods is *u* and *v*, the number of time periods is *T* and the efficiency score measures are conditional upon the choice, S^{s}, of reference production possibility set.[3] The frontoer technology s is commonly made to change over time, using *s* = *u* or *v*, when calculating productivity change for consecutive periods. The Malmquist index is calibrated as in Caves et al. such that *u* < *v*, implying that the number is greater than 1 for progress and less than 1 for regress.

The definitions of the input- and output oriented Farrell efficiency indices appearing in (2) are:

$$\begin{aligned}&E_{1,it}^{s}(x_{it},y_{it};S^{s})=Min\ \left\{\theta\middle|(\theta x_{it},y_{it})\in S^{s}\right\}\\&E_{2,it}^{s}(x_{it},y_{it};S^{s})=Min\left\{1/\phi\middle|(x_{it},\phi y_{it})\in S^{s}\right\}\ ,t=1,..,T,i=1,...,J\end{aligned} \qquad (3)$$

The efficiency measures are calculated relative to the benchmark frontier *s*. This implies that the efficiency scores are not bounded to be less than or equal to one, but may be greater if the period *t* observations are outside the benchmark frontier. The DEA estimates of the efficiency scores,

[3] For simplification the vectors of inputs and outputs for the two periods *u* and *v* being arguments in the index function is represented just by *u* and *v* on the left-hand side of (2).

$\hat{E}^{s}_{d,it}(d=1,2)$, are calculated by setting up the standard linear programming problems of DEA in the case of using a benchmark production possibility set *s* to define the frontier.[4]

The productivity interpretation of the Malmquist index follows from noting that the definition of the efficiency measures involved implies that observed productivity is compared with maximal productivity at the border of the technology set, keeping either inputs or outputs fixed. The Malmquist index captures the relative change in efficiency for two periods, and since the reference frontier is the same this relative measure has the interpretation of productivity change. In order to calculate the efficiency scores the data set must include time periods. This is straightforward: the observations used to support the frontier indexed *s* must be specified (in the literature observations for a specific year are often used), 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*.

Productivity as measured by the Malmquist index, based on the production possibility set (1) above, may be influenced by changes in the scale of the operation, but two units that have the same ratio of outputs to inputs should be viewed as equally productive, regardless of the scale of production (Grifell-Tatjé and Lovell, 1995). Doubling all inputs and outputs keeping input and output mixes constant does not change productivity, even though the technology S^{s} has variable returns to scale (VRS). The relevant reference set for measurement of total factor productivity (TFP) is therefore one that is homogenous of degree 1 in the input-output vector, and the homogenous set that fits closest to the technology is the envelopment of S^{s} defined by

$$\overline{S}^{s}=\left\{(x,y)\middle|(\gamma x,\gamma y)\in S^{s}\right\} \qquad (4)$$

where γ is a free positive scalar. Only if the underlying technology exhibits constant returns to scale will the sets defined in (1) and (4) be equal. This envelopment is often termed the CRS frontier, even though the underlying technology is VRS, and is illustrated in Figure 1 by the ray from the origin labelled CRS.

[4] If a specific year is chosen as the basis for the reference technology the variable returns to scale (VRS) specification will not yield feasible solutions for observations having smaller inputs than the smallest observed for the reference set.

The homogenous envelopment can be used to define the concept of technically optimal scale (Frisch, 1965). This is the scale where the returns to scale is one, and is illustrated in Figure 1 as the tangent point P^{tops} of the CRS line and the VRS frontier.[5] A proper TFP measure is thus obtained by only using the technology information about changes in the technically optimal scale over time. Such a subset of the technology set, termed TOPS in (Førsund and Hjalmarsson, 2004b) is defined by:

$$TOPS^s = \left\{(x,y) \middle| \ \varepsilon(x,y) = 1, (x,y) \in S^s\right\} \qquad (5)$$

where $\varepsilon(x,y)$ is the scale-elasticity function. From classical production theory we know that the productivity is maximal at optimal scale where returns to scale, ε, is one, thus this is a natural reference for productivity changes over time. The TOPS set is also the intersection of the efficient boundary of the technology S^s (the frontier production function) and the boundary of its envelopment $\overline{S}^s$.

An illustration in the two-period case is provided in Figure 1. Observations of the same unit are indicated by P_1 and P_2. The two corresponding VRS frontiers are drawn showing an outward shift indicating technological progress. In Figure 1 the TOPS point for period 2 is labelled P^{tops}. Just as the productivity should be unchanged if the input-output vector is proportionately scaled, a measure of productivity should double if outputs are doubled and inputs are kept constant, and vice versa. A productivity measure should therefore be homogenous of degree 1 in outputs and of degree (-1) in inputs. A measure of productivity change over time, based on all the inputs and outputs for the unit in question for two periods, should similarly be homogeneous of degree 1 in outputs from the last period and inputs from the first period and homogeneous of degree (-1) in outputs from the first period and in inputs from the second period; i.e., if outputs in the second period or inputs in the first period doubles the TFP measure should also double, and if outputs in the first period or inputs in the second period doubles then the TFP measure should be reduced to one half. Using the subset TOPS is one way of obtaining the required homogeneity properties of a Malmquist productivity index.

[5] 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 (Førsund and Hjalmarsson, 2004a).

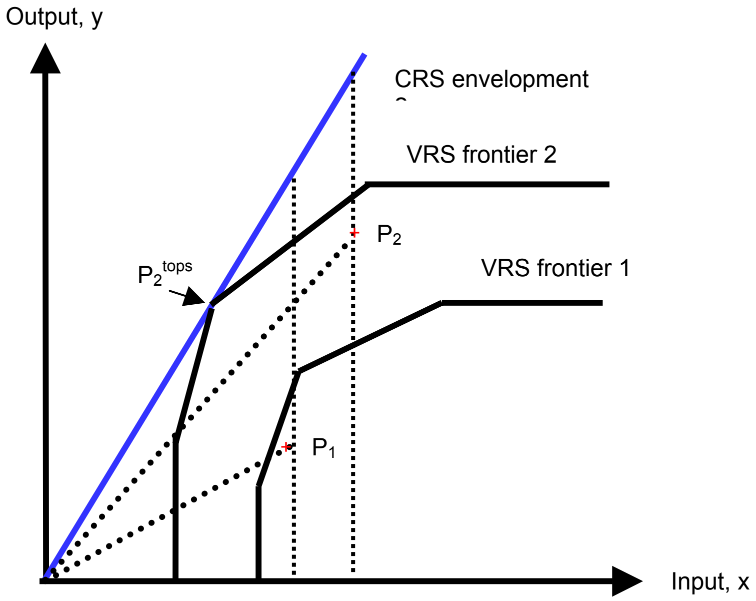


Figure 1. The Malmquist productivity index

A further question is whether to use the envelopment of the technology of a single year or several years as the reference for the productivity index. One consideration is whether the resulting productivity measure is *circular* (Berg et al, 1992). We will be interested in pointing to years with strong or weak productivity growth, so we need circularity in order to interpret the results in such a way (Gini, 1931). We will use as a reference technology a sort of average technology by using the envelopment of all technology frontiers as a fixed reference frontier, i.e. $S^s = \bigcup_t S^t$, thereby fulfilling the circularity condition while at the same time utilising technology information from all time periods. In Tulkens and van den Eeckaut (1995) this type of frontier was termed the *intertemporal* frontier.[6] As is common with indices, performance is calculated using information that may not have been available in the first period, but this is consistent with retrospective evaluation.

Using a linear homogeneous envelopment implies that the orientation of the efficiency index does not matter. The estimator of the Malmquist index then simplifies to:

[6] In Pastor and Lovell (2005), missing out on the reference to Tulkens and van den Eeckaut, it is called the global frontier.

$$\hat{M}_i^s(u,v) = \frac{\hat{E}_i^s(x_{iv}, y_{iv}; \hat{\hat{S}}^s)}{\hat{E}_i^s(x_{iu}, y_{iu}; \hat{\hat{S}}^s)} \quad , \quad i = 1,..,J,\ u,v = 1,..,T, u \neq v \tag{6}$$

where superscript *s* here now indicates that all data is used as the technology reference set. The Malmquist productivity estimator is conditional on the estimator, $\hat{\hat{S}}^s$, for the linear homogeneous envelopment set in (4).

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 and the frontier rests on outlier observations. 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 efficiency scores are correspondingly optimistically biased. The sampling bias for a given observation can be expected to be higher the lower the number of other observations in the sample. Banker (1993) proved in the one input – one output case that as the number of observations goes towards infinity, the distance between the DEA estimate and the true efficiency score goes towards zero, i.e. the DEA estimator is consistent. In Simar and Wilson (2000) generalisations to multiple outputs and inputs are reviewed. 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. Following Simar and Wilson (1998), the data generating process (DGP) assumes in general that the sample observations (x_i,y_i) are realisations of independent identically distributed variables on the production possibility set with a probability density function. In our setting with, e.g., a radial output-oriented efficiency variable E_2 distributed on (0,1] we assume that the observations are generated by randomly drawn efficiencies from the true efficiency distribution, with exogenously given output levels and input 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, i.e. the efficiency distribution is independent of output scale and input mix, but this can be relaxed with a more complicated DEA bootstrap methodology (Simar and Wilson, 2000).

Bootstrapping is a way of testing the reliability of the dataset, and works by creating pseudo replicate data sets using resampling (Efron, 1979). The resampling is done on the basis of the calculation of efficiency scores relative to the VRS frontier for each time period. Kernel density estimation (KDE) is used to smooth the empirical distribution of the original efficiency scores, using reflection (Silverman, 1986), in order to avoid the accumulation of efficiency score values of 1. This is necessary in order to have a consistent estimator of the efficiency score distribution at the boundary of the distribution where the efficiency score is 1. The pseudo observations are then created by projecting all inefficient observations to the original DEA frontier, and drawing randomly an efficiency score for each unit (including the originally efficient ones) from the KDE distribution. When, as in our case, the inefficiency is assumed to be output-oriented, the level of each output *m* is calculated as:

$$y_{imt}^{ps} = \frac{y_{imt}}{\hat{E}_{2it}^{s}} E_{2t}^{KDE}, i = 1,..,J\,,\ m = 1,...,M\,,\ t = 1,...,T \tag{7}$$

where E_{2t}^{KDE} is a draw of the KDE distribution for the efficiency score. A new DEA frontier is then estimated on these pseudo observations (x_i, y_i^{ps}), each generated by mimicking the original DGP, as if the original DEA estimated frontier were the true frontier. 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. Now, going back to each run for period *t* the Malmquist productivity index, given by (6), is calculated with reference to the linear homogeneous benchmark technology created for the pooled set of all pseudo observations.

Figure 1 may illustrate the procedure for observations P$_1$ and P$_2$. These observations may be regarded as pseudo observations created as explained above. The CRS envelope may be regarded as the envelope created on the full pseudo observation set. The value of the Malmquist indices are found by using the points on the CRS envelopment following the vertical lines from the pseudo observations.

Bias correction in DEA using bootstrapping, following Simar and Wilson (1999), is based on an assumption by analogy on the distributions of the estimators, implying that the difference between the Malmquist estimator based on pseudo data and the DEA-based estimator is distributed like the distribution between the DEA estimator and the true Malmquist index, assuming estimators to be consistent:

$$(\tilde{M}^s(u,v) - \hat{M}^s(u,v))\Big|\hat{S}^s \sim (\hat{M}^s(u,v) - M^s(u,v))\Big|S^s\,,\ u,v = 1,..,T, u \neq v \tag{8}$$

Here M^s is the true unknown efficiency, $\hat{M}^s$ is the original DEA estimate, $\tilde{M}^s$ is the bootstrapped estimate and S^s and $\hat{S}^s$ are the theoretical production possibility set and its DEA estimate, respectively.

Building upon (8) the bias of the Malmquist productivity estimator can be estimated. However, it is pointed out in Simar and Wilson (2000) that the bias correction may create additional noise in the sense that the mean square error of the bias-corrected score may be greater than the mean square error of the uncorrected estimator. In our case the bias-corrected values are quite close to the original estimates for the Malmquist index, as is commonly the case for bootstrapping the Malmquist index (Edvardsen et al., 2006). Three units have relatively large differences between the original estimate and the bias-corrected, while most other units have small differences that are both positive and negative. We have used the bias-corrected estimates.

This problem motivated Simar and Wilson (1999) to suggest a direct way to calculate the confidence intervals so that they could be centred around the original DEA estimates rather than the bias-corrected estimates if the former had a lower mean square error. The confidence interval limits may be defined by

$$\Pr(-\hat{b}_{\alpha i} \leq \tilde{M}_i^s(u,v) - \hat{M}_i^s(u,v) \leq -\hat{a}_{\alpha i}\Big|\hat{\hat{S}}^s) = 1-\alpha,\ i = 1,..,J,\ u,v = 1,..,T, u \neq v \tag{9}$$

The estimates for the limits are found from the distribution of $(\tilde{M}_{ib}^s(u,v) - \hat{M}_i^s(u,v))$ for *b* = 1,..,*B* by sorting in increasing order and finding the values matching the chosen degree of confidence. The estimated (1 - α) confidence interval for the true efficiency score $M_i^s(u,v)$ is then

$$\hat{M}_i^s(u,v)+\hat{a}_{\alpha i} \le M_i^s(u,v) \le \hat{M}_i^s(u,v)+\hat{b}_{\alpha i},\ i=1,..,J,u,v=1,..,T,u \ne v \qquad (10)$$

3. Data

The Directorate of Taxes is responsible for assessing taxes and for collecting them. The local tax offices in Norway use about 60 % of all labour of the Directorate, and are responsible for tax assessment for all types of income tax. In that connection the tax offices are also responsible for keeping track of changing addresses of persons and companies. A motive for collecting primary statistics at the level of a local tax office is then that an updated address register of people and firms is necessary for the quality of tax assessment. Such statistics are also collected to help other public sectors. Collecting data on outputs makes it possible to keep track of the work load of a tax office by the central decision-making unit. This is necessary in order to obtain a realistic picture of the local activities and control the allocation of resources to offices.

The present study is restricted to use pre-existing data. In view of the observation made in the introduction of difficulties with measuring inputs and outputs in the public sector since it is not operating through markets, it is pertinent to ask if the available data are good enough for the purpose of measuring efficiency. The Directorate has answered cautiously affirmative since statistics of the main activities in the form of many detailed indicators are kept for internal use, and the Directorate of Taxes has had an extensive discussion about the most relevant measures for outputs and inputs. Furthermore, the data set has been controlled in several different ways, e.g. finding extreme values by inspecting the distribution of variables and partial productivities, abnormal changes from year to year, etc., and should have ensured an acceptable quality of the data.

Although the data are not collected primarily to serve the purpose of efficiency and productivity studies of offices the existing output data is not based on input costs, but constitute independent quantity measurements and thus may be used for such studies. One could wish more information on quality both of outputs and inputs, but since efficiency and productivity studies of the nature reported here and in Førsund et al. (2006a,b) are new to the Directorate, the effort of gathering relevant variables for the study based on existing data was seen as enough for a first step.

The list of the variables chosen for the study together with some key information about the variables is given in Table 1. Only one input is specified; the total use of labour measured in man-years, net of labour used for administration.[7,8] Six outputs are specified representing the main activity areas. The main activities are to process tax returns from individuals and returns from the two types of businesses that are specified; self-employed and limited companies. In

Table 1. The data

Variable	Year	Range	Mean	Std. Deviation
X: Number of man-years net of administration	2002	378	70	102
	2003	300	69	102
	2004	377	67	102
Y1: Number of people relocated during the year registered by home address and number of immigrations and emigrations.	2002	96 365	6 153	10 686
	2003	100 523	6 243	11 127
	2004	109 693	6 562	11 973
Y2: Number of false registrations detected by control activities.	2002	799	39	97
	2003	1 526	48	156
	2004	3 299	70	337
Y3: Number of tax returns from employees and pensioners	2002	413 424	34 540	46 818
	2003	416 511	35 226	47 531
	2004	423 221	35 334	48 015
Y4: Number of complaints on tax assessment	2002	16 255	647	1 839
	2003	10 009	537	1 211
	2004	11 169	497	1 245
Y5: Number of returns from non-incorporated businesses.	2002	31 709	3 230	3 411
	2003	32 871	3 318	3 522
	2004	33 931	3 302	3 669
Y6: Number of corporate tax returns.	2002	33 038	1 624	3 484
	2003	31 022	1 632	3 304
	2004	31 194	1 655	3 338

[7] The netting is based on detailed time-use studies.

[8] Labour is the dominating input, counting for about 80% of total costs, as is also the case for Belgian tax offices reported in Moesen and Persoons (2002).

addition one variable covering treatment of complaints, and two variables covering activities checking the information about addresses are included.

For the purposes of testing variables and estimating the frontier reference for the productivity changes, we have chosen to pool the data for the 98 units for the three years for which we have observations since it seems reasonable to assume that the technology is stationary over the three-year period.

Although the choice of outputs were made internally at the Directorate, the relevance of two of the variables, number of false registrations detected (y_2) and number of complaints on tax assessment (y_4), were questioned. We have therefore carried out a stepwise test procedure using bootstrapping to test whether the addition of these variables made a significant change of efficiency scores. Starting first with including y_4, but keeping y_2 out, it turned out that this variable made a significant impact, and then introducing y_2 this was also significant, although just so. We have therefore kept the specification shown in Table 1.

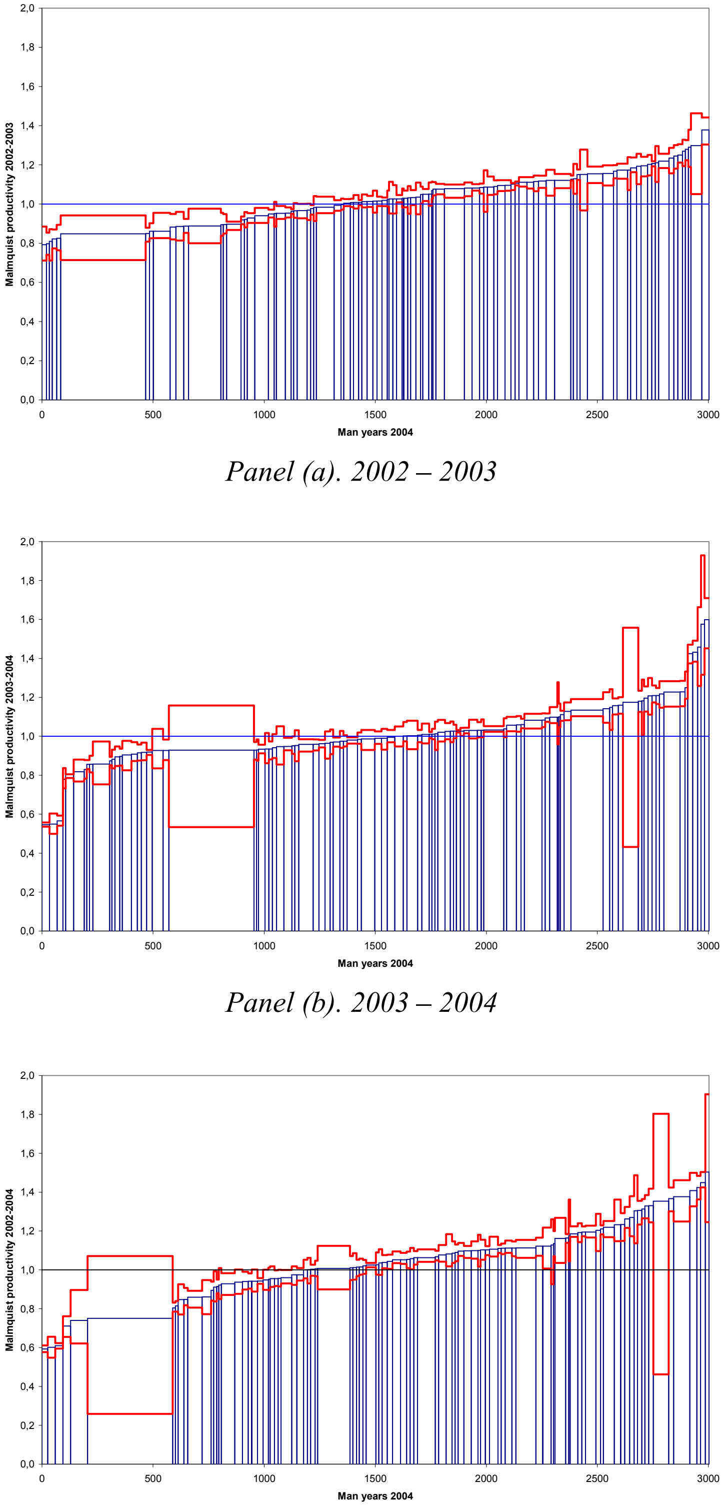
4. Productivity development

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 Eq. (6) 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. We have assumed that the true values of the Malmquist index are independent over time and have followed the bootstrap procedure outlined in Section 2 (Simar and Wilson, 1999), but without assuming correlation of the Malmquist indices over time. As reported in Simar and Wilson (1998) the mean-square errors may increase using the bias correction and increase so much that it is better to use the original DEA estimates. However, in our case checking the test statistic provided in Simar and Wilson, we have chosen to use the bias-corrected Malmquist indices in the illustrations below. Confidence intervals are established following the procedure outlined in Section 2.

The productivity development for the units between 2002 -2003, 2003-2004 and 2002 - 2004 are set out in panels (a-c) in Figure 2. All the 98 tax offices are shown, represented by histograms with the width proportional to man-years (for 2004). The histograms are sorted for ascending values of the bias-corrected productivity index. In addition the limits of the lower and upper 95 % confidence interval are shown. The horizontal line at the value of 1 delimitates units with productivity decline and increase.

In Panel (a) the development from 2002 to 2003 shows that units representing about 46% of the labour (in 2004) have had a productivity decline, starting at an index value of 0.80, meaning that the productivity has declined with 20% for this tax office, while the other half have had a productivity increase up to 38%. The group of offices with the highest decline and the highest increase are both on the small side. The largest units and some medium-sized offices show productivity decline, while most medium-sized units show increase. The confidence intervals show that the large and medium-sized units have the widest intervals. Both the groups with the highest productivity decline and the strongest increase have the narrowest intervals (with some exceptions), implying that their productivity developments are rather accurately estimated. A strategic question is whether decline and increase are significant. This can be tested simply by inspecting whether the value of 1 is contained in the confidence interval. We see that all the large and medium-sized units with decline in productivity have had significant declines, while for some medium-sized units with productivity increase this is not significant.

The development shown in Panel (b) for 2003 – 2004 reveals that more units have productivity decline distributed over a wider interval than the previous period, starting at 45% regress, but it is a small tail of small units that has these values, then productivity decline is at the same level as the previous period. The units with decline now represent 54% of labour (in 2004). The productivity growth is weaker with most of the units achieving less growth than 20 %. There is a marked tail of units on the smaller side with growth in the range of 42 - 55%. But units in this



Panel (a). 2002 – 2003

Panel (b). 2003 – 2004

Panel (c). 2002 – 2004

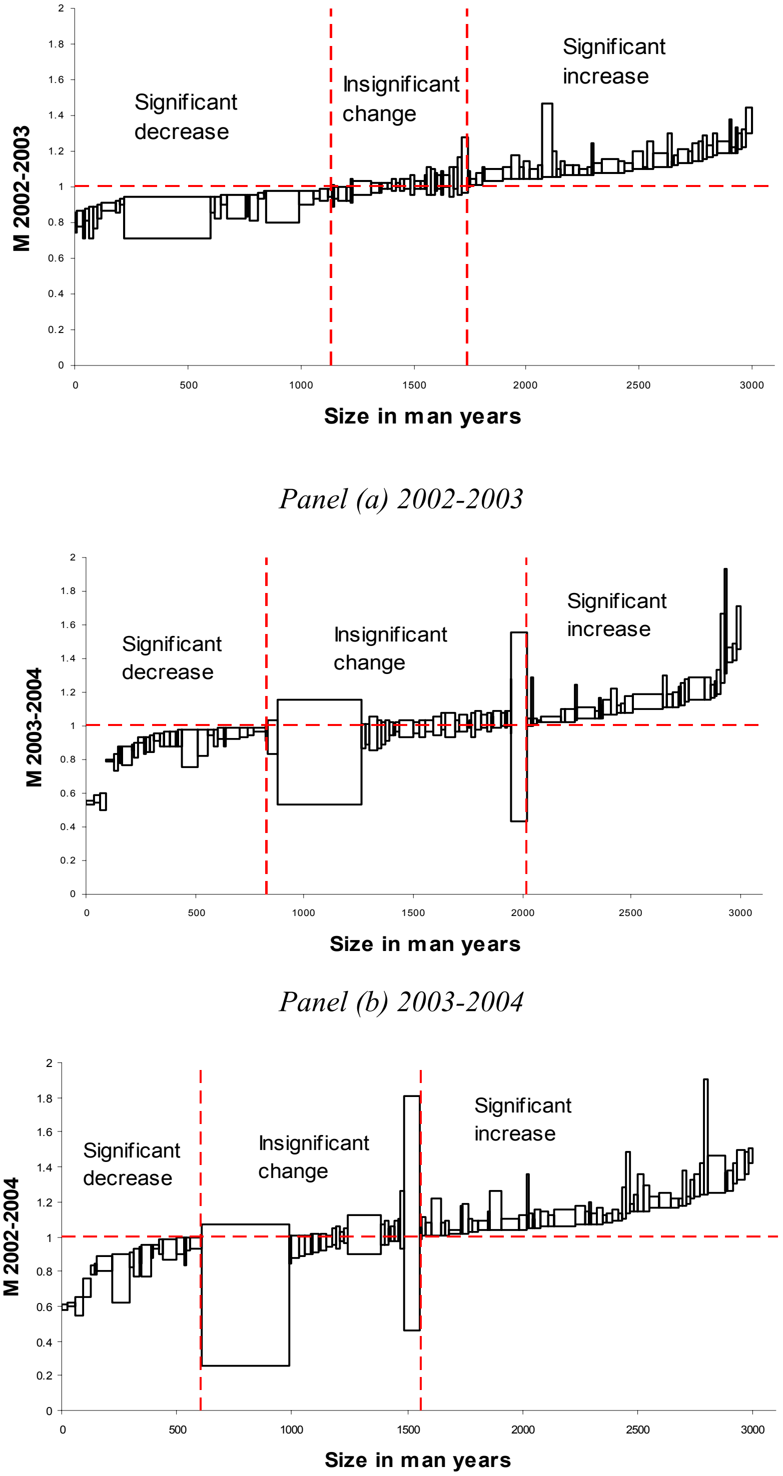
Figure 2. Malmquist productivity index estimates and confidence intervals. Width of histograms is proportional to labour input

tail tend to have wider confidence intervals. The two largest units have increased their index values, and the second-largest has even moved into productivity progress.

The development over the whole period is set out in Panel (c). We can clearly see that the development of the sizes of the different groups gives us a kind of average picture for the two separate periods, so if one is interested in the total picture this is the useful diagram. The higher confidence or the larger units revealed for the period 2003 – 2004 is also evident for the total period. However, treating the two periods separately reveals more details, and it is interesting to see the difference between the periods in view of the reorganisation plans. The productivity change distribution for the bias-corrected Malmquist index ranges from a 40% decline to a 45% increase. Taken at face value the results indicate that units representing 39% of the man-years (in 2004) have had a productivity decline over the three years, while units representing 61% have had a productivity improvement. Among units with productivity improvement the small ones dominate. Some average sized units have had slight improvements while others have experiences decline.

Testing hypothesis whether an office has had a significant decline or increase in productivity, as stated in Simar and Wilson (1999) is the benefit of bootstrapping. Instead of commenting further on the confidence intervals shown in Figure 2 we have in Figure 3, in order to illustrate the testing, set out special *Edvardsen significance diagrams* focussing only on the confidence intervals for the units. The units are grouped in three groups, units with significant decrease in productivity, units with insignificant productivity change and units with significant increase. In the first group the units are sorted according to ascending values of the upper limit of the confidence interval, in the second group the unit are sorted according to ascending values of the mid value of the confidence interval[9], and in the third group the units are sorted according to ascending values of the lower limit of the confidence interval. Using the mid value of the confidence interval as a sorting value illustrates the position of the interval relative to the crucial value of 1 signifying no productivity change. The size of the boxes is based on the same relative share of total man-years (for 2004) as in Figure 2.

[9] This value is equal to the bias-corrected estimates of the Malmquist index (not shown in the diagram).



Panel (a) 2002-2003

Panel (b) 2003-2004

Panel (c) 2002-2004

Figure 3. Significance testing: units grouped by the nature of the significance of productivity change. Sorted by lower limit, mid point, and upper limit of confidence interval respectively. Width of boxes proportional to labour input.

In Panel (a) we have that in the group having significant productivity decrease there are 25 units representing 38% of the man-years. The number of units and the change in shares of labour is set out in Table 2. There is a relative overrepresentation of large units in the group with significant decrease. The group of insignificant change of 28 units represents 19% of the labour input. The four first units in the group are very close to ending up in the group of units with significant decrease, while the right-hand tail have units being very close to having a significant increase. There is an overrepresentation of small units in the group with insignificant change. The group of significant increase have the largest number of 45 units representing 43% of total labour, and the average-sized units dominate the group with significant increase in productivity.

Panel (b) reveals a structural change from the period 2002-2003 to the period 2003-2004. The group with significant decrease has increased in number from 25 to 30. However, the share of labour has decreased markedly from 38% to 27%, implying a substantial reduction of average size of units in this group. Some large units have moved to the group of insignificant change, and their confidence intervals have widened markedly. Both the number of units and the share of labour have increased in this group; the number from 28 to 33 and the share from 19 to 40%. The average size of units in the group has thus increased considerably. Both the share of labour and the numbers have been reduced markedly in the group with significant increase in productivity; from 43% to 33 %, and from 45 to 34 for shares and units, respectively, but the relative average size has been kept about the same. The shrinking of both the groups with significant decline and significant decrease, and the increase in the group of insignificant change, reflects the general tendency of a widening of the confidence intervals.

Table 2. Productivity change in percent of total man-years 2004.
Number of units in parenthesis

Periods	**Productivity decrease**		**No significant change**	**Productivity increase**	
	Bias corrected	Significant		Bias corrected	Significant
2002-2003	55	38 (25)	19 (28)	45	43 (45)
2003-2004	47	27 (30)	40 (33)	53	33 (34)
2002-2004	39	20 (20)	32 (24)	61	48 (54)

Panel (c) showing the overall 2002-2004 development is clearly more like the picture shown in Panel (b) in terms of the size of confidence intervals, which are typically wider, especially for large units, than in Panel (a). The group having significant decrease now consists of units closer to average size, having 20 units and a share of 20% of labour. The confidence intervals for the large units are especially wide for the group with insignificant change, and the group has 24 units representing 32% of the labour. The group of units with significant increase in productivity has increased both in shares and number; to 48% and 54 respectively, indicating a somewhat smaller average size than total average. Comparing this group with the groups shown in the two other panels we have that the period 2003-2004 had a relative setback in number of units with significant productivity increase, but that over the period as a whole some units not having significant increase neither in the period 2002-2003 nor in the period 2003-2004 experienced significant increase over the total period. The group having significant decrease is smaller than for both the separate periods, and the group having significant increase is larger than for the two separate periods.

More reduced data densities in the neighbourhood of large units make the determination of the productivity score more uncertain. The implication is that we can trust more the results for the small units, but that we must be careful when using productivity figures for the large units. This is especially evident for the largest unit in Panel (c) of Figure 2. But note that since we measure the indices relative to a CRS frontier we cannot say that size as such is an explanation for wide confidence intervals. Units can stand apart because of the nature of the output mix, such as the relationship between tax returns form persons and from firms, and such features may be correlated with size.

Comparing the change in the resources used and the productivity scores provides a further characterization of the nature of productivity growth (see Førsund and Kalhagen (1999), Førsund et al., 2006a). In Figure 4 productivity change from 2002 to 2004 is shown together with the relative change in use of man-years. The relative area of a circle is proportional to man-years in 2004. The horizontal axis measuring labour change is placed at the level of 1 for productivity change, while the vertical axis measuring productivity change is placed at zero change of labour use. To the left of the origin use of man-years have decreased from 2002 to 2004 while to the

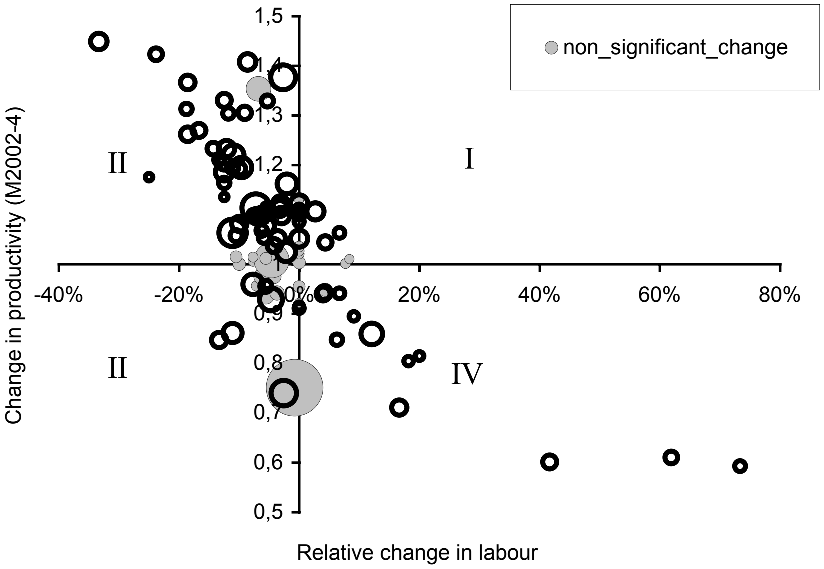


Figure 4. Productivity and labour change 2002-2004.
Size of circles is proportional to labour

right the use has increased. The total range is from -40% to +80%. Together with the horizontal axis at 1 delimitating the units with productivity decrease and increase, respectively, the vertical axis through zero change in labour form four quadrants numbered I to IV. In Quadrant I units have had both productivity growth and increase in man-years. Such units may be said to have experienced *efficient labour increase*. This quadrant has the fewest units. The unit with the highest labour increase of 8.3% has had a productivity growth of 1% which is the highest in Quadrant I, while the unit with the highest productivity growth of 11% has had an increase of 2.7% in labour. The units in Quadrant II have also had productivity growth, but experienced labour reductions. This may be termed *efficient labour savings*. This quadrant has the highest number of units. The unit with the highest productivity change has had an increase of 45% (maximal of all units) and reduced labour with 33% (also maximal). In quadrant III productivity decrease is combined with labour decrease. This is *inefficient labour savings*. There are relatively few units in this quadrant. The unit with the highest labour decrease has had a productivity decrease of 15% and a labour decrease of 13.3%, while the unit with the highest productivity decline reduced the productivity with 26% and labour with 2.6%. Units in Quadrant

IV have the worst of both worlds with decreasing productivity and increasing costs. This is *inefficient labour increase*. Three units with almost the same maximal productivity decline, 40%, have had labour increases ranging from 42% to the maximal increase of 73%.

A few units are extreme in their change in labour, like the three units in Quadrant IV and also three units in Quadrant II. This may be explained by reorganisation and moving of tasks between offices. The rest of the units are found in the interval -20% to +20% change in labour. But is revealed that a relatively large group of units have no change in labour, with a dispersion from a decrease in productivity of 9% to an increase of 12%. Most of the other units belong either to Quadrant II with efficient labour decrease, or Quadrant III with inefficient labour decrease, but there is still a clear negative correlation between productivity and labour change.

To see this more clearly the units with insignificant productivity change (i.e. the middle group shown in Panel (c) in Figure 3), are shown in Figure 4 with grey filling of the circles. It must be the case that units with significant productivity increase are found in Quadrants I and II with positive growth, and units with significant decrease in productivity are found in Quadrants III and IV with negative growth. The largest unit with insignificant productivity decrease belongs to Quadrant III, while the second-largest unit with insignificant productivity increase belongs to Quadrant II. It is interesting to note the relatively large productivity decline of the largest unit, 25%, and the high productivity increase of 35% of the relatively large unit, but still both have insignificant change (the units are clearly exposed in Panel (c) in Figure 3). There are only six units with significant productivity decline in Quadrant III with inefficient labour saving. Only three units appear in Quadrant I with efficient labour increase and with significant productivity growth, and 11 units are in Quadrant IV with inefficient labour increase and significant productivity decrease.

Total productivity change

Concerning average growth rates measured by the Malmquist index we will use two variants of a bottom - up approach. One approach, linked to Farrell's way of measuring how the mean performance of a sector is compared with the frontier, is to form an average tax office by averaging inputs and outputs and then enter this unit as a micro unit in the calculations. Another

Table 3. Average growth rates in percentage

Period	Growth measure	Original point estimate	Bias-corrected
2002-2003	Average unit	2.4	2.5
	Mean	4.5	4.5
2003-2004	Average unit	5.3	2.0
	Mean	3.2	2.3
2002-2004	Average unit	7.8	4.5
	Mean	7.3	6.2

more conventional approach is to take some mean of the individual results. We have done both approaches and the results are set out in Table 3. We have chosen to use the simple arithmetic mean. The difference in aggregated results between original point estimates of the Malmquist indices and the bias-corrected ones are also shown.

For the first period 2002 – 2003 the results for the two bottom-up measures are very similar in the for the two types of estimates of productivity, showing a 2.5% growth for the average unit-measure and 4.5 % for the mean value of the individual estimates. For the second period, however, the estimates differ, especially for the average unit measure being 5.3% for the original estimates and 2.0% for the bias-corrected estimates, and 3.2% and 2.3% for the mean results, respectively. The total period 2002 – 2004 also reveals some differences for the average unit measure, but relatively smaller difference for the mean measure showing 7.3% and 6.3%, respectively. The differences between the mean values of the original estimates and the bias-corrected means for the periods 2003 – 2004 and 2002 -2004 show that negative bias corrections dominate the bias-corrected measures.

5. Conclusions

Productivity measurement in the public sector may be based on a top-down approach or a bottom-up approach. The advantage of a bottom-up approach followed here is that existing primary-data collection at the micro level of parallel service production units can be utilised, not only to measure the aggregate productivity performance of interest for external use, like reporting the figure of 6 % productivity increase on the average over the three-year period, but also to reveal the productivity performance of individual units. The data can thus give necessary information for providing explanations for differences in productivity performance across micro units.

The main objective of performance measurement of units is to present results in ways that facilitate improvement of performance. This is of special importance for a public service production sector not selling the services in a market and facing accountability and stakeholder interest in performance. The present study has shown that is of crucial importance to use methods that enables us to make a statistical assessment of the uncertainty of productivity estimates that are the “engine” of performance measurement over time. The results show that large units would have appeared to have a better productivity performance when uncertainty is not accounted for than they get with explicit treatment of uncertainty. Establishing confidence intervals for productivity performance makes it possible to test hypotheses about declining or increasing productivity in a rigorous way. The share of labour used in units with productivity increase decreased from 58% to 40%, and the share with apparent productivity decline decreased from 42% to 17% when looking at the significant changes only.

The productivity results reveal changes even over short periods. Part of the changes must be attributed to internal budgeting procedures naturally lagging real changes in tasks that are mainly exogenous. As more time periods accumulate productivity analyses should provide more valid information on inherent qualities of tax offices expressed by the labels efficient labour increase, efficient labour savings, inefficient labour savings and inefficient labour increase used in this study. Taken at face value units representing 42% of the man-years have had productivity decline and 58% have had productivity increase within the range of -40% to +45%, respectively, and resulting in an overall productivity increase of about 6%. 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 wide, while they are narrow for small units. It is also of interest to note that both small and large tax offices are found in both the two groups of offices with significant decline and increase of productivity respectively. Therefore causes of productivity differences cannot be attributed to size in general, but may be sue to product mix.

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. This is left for further research. A good start will be to study the units appearing as the units with the best productivity performance in Figures 2 - 4, and check, e.g., their pattern of use of resources and composition of outputs compared with the average in order to generate hypotheses about factors explaining productivity differences.

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.

Berg, S. A., F. R. Førsund, and E. S. Jansen, (1992): "Malmquist Indices of Productivity Growth during the Deregulation of Norwegian Banking, 1980-89," *The Scandinavian Journal of Economics* 94, Supplement. Proceedings of a Symposium on Productivity Concepts and Measurement Problems: Welfare, Quality and Productivity in the Service Industries, S211-S228.

Bird, S. M., Sir D. Cox, V. T. Farewell, H. Goldstein, T. Holt and P. C. Smith, (2005): "Performance indicators: good, bad, and ugly," *Journal of the Royal Statistical Society*, Series A, 168 (Part 1), 1-27.

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.

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.

Edvardsen, D. F., F. R. Førsund, W. Hansen, S. A. C. Kittelsen, and T. Neurauter (2006): "Productivity and regulatory reform of Norwegian electricity distribution utilities," in T. Coelli and D. Lawrence (eds), *Performance measurement and regulation of network utilities*. Edward Elgar Publishing Company, 97-131.

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. (1997): “The Malmquist productivity index, TFP and scale,” Memorandum No 233, Department of Economics, Göteborg University.

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

Førsund F. R. and L. Hjalmarsson (2004b): “Calculating scale elasticity in DEA models,” *Journal of the Operational Research Society* 55, 1023-1038.

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.

Førsund, F. R., S. A. C. Kittelsen, and F. Lindseth (2005): “Efficiency and productivity of Norwegian tax offices,” *Memorandum* 29/2005 from the Department of Economics, University of Oslo.

Førsund, F. R., S. A. C. Kittelsen, F. Lindseth and D. F. Edvardsen (2006a): “The tax man cometh - but is he efficient?” *National Institute Economic Review* 2006 197(July), 106-119.

Førsund, F. R., S. A. C. Kittelsen, and F. Lindseth (2006b): “Productivity of tax offices in Norway”, paper presented at the 31st CEIES Seminar - Are we measuring productivity correctly?, Rome 2006, pp. 149-169.

Gini, C. (1931): “On the circular test of index numbers,” *Metron* 9 (2), 3-24.

Grifell-Tatjé, E. and C. A. K. Lovell (1995). “A note on the Malmquist Productivity Index,” *Economics Letters* 47, 169-175.

Malmquist, S. (1953): “Index numbers and indifference surfaces,” *Trabajos de Estadistica*, 4, 209-224.

Moesen, W. and A. Persoons (2002): “Measuring and explaining the productive efficiency of tax offices: a non-parametric best-practice frontier approach,” *Tijdschrift voor Economie en Management* XLVII(3), 399-416.

Pastor, J. T. and C. A. K. Lovell (2005): “A global Malmquist productivity index,” *Economics Letters* 88, 266-271.

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 (1999): “Estimating and bootstrapping Malmquist indices,” *European Journal of Operations Research* 115(3), 459-471.

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.

Tulkens, H. and P. van den Eeckaut (1995): “Non-parametric efficiency, progress, and regress measures for panel data: methodological aspects,” *European Journal of Operational Research* 80, 474-499.