J Prod Anal (2015) 43:269–279
DOI 10.1007/s11123-015-0435-1

Productivity of tax offices in Norway

Finn R. Førsund · Dag Fjeld Edvardsen · Sverre A. C. Kittelsen

Published online: 15 February 2015

Abstract The performance of local tax offices is studied over time using data envelopment analysis to calculate Malmquist productivity indices. The index has the proper homogeneity properties of a total factor productivity index. One input, cost, and six output categories of the main service activities carried out by tax offices, are specified. A bootstrap approach is applied to establish confidence intervals for the individual indices enabling an identification of units that have significant productivity decline, growth, or no change. A novel visual test groups units into these three possible categories. This way of showing consequences of uncertainty should facilitate more tailor-made policies to promote efficiency and productivity improvements. Productivity changes are distributed from a 26 % decline to a 35 % increase over the three-year period with an average growth of 4 %. Inspecting individual unit results, the confidence intervals tend to be wider the larger the units, thus providing more accurate insights than point estimates for actions to improve productivity. Looking at positive and negative changes in cost and productivity together the development of offices is classified into four categories of interest to policymakers; efficient cost increase, efficient cost saving, inefficient cost saving, and inefficient cost increase.

The paper is written within a project undertaken by the Frisch Centre for the Norwegian Directorate of Taxes to evaluate the impact on efficiency and productivity of a recent reorganisation of the Directorate. We are indebted to Frode Lindseth for preparing the data and giving valuable insights and comments.

F. R. Førsund (✉)
Department of Economics, University of Oslo, Oslo, Norway
e-mail: finn.forsund@econ.uio.no

F. R. Førsund · S. A. C. Kittelsen
The Frisch Centre, Oslo, Norway

D. F. Edvardsen
Catenda, Oslo, Norway

Keywords Tax office · Malmquist productivity index · DEA · Bootstrap · Confidence intervals

JEL classification C60 · D24 · H21 · L89

1 Introduction

In many countries there has been an increased emphasis in recent years on promoting accountability of use of resources in the public sector in order to promote efficiency and productivity. A necessary first step is then to measure efficiency and productivity. Tools for doing this have been developed for the situation when there are a sufficient number of providers of the same service. The purpose of this study is to estimate the productivity development of tax offices of Norway. The productivity development is analysed using the Malmquist productivity index (Caves et al. 1982) which builds upon efficiency scores calculated using the non-parametric data envelopment analysis (DEA) for a rather short time period; 2002–2004. A reorganisation of the tax offices started in 2005, forming five tax regions where tax offices share tasks for the entire country independent of localisation. It is then not possible to get data for individual offices after 2004. In addition, all older data have to be constructed based on existing different types of data and expert knowledge. Therefore 2002 was the first year for which variables suitable for productivity analysis could be established taking available in-house resources for this job into account. The purpose of studying past productivity is to have a benchmark for measures of productivity development after the reorganisation and the period 2002–2004 was deemed to be sufficient by the Tax Directorate. However, in view of the short time period we would like to emphasis the development of the tools for visualising individual productivity developments and their confidence intervals as a contribution of the paper.

An important issue for policy implications of evaluations of productivity performance is whether changes are significant or not. A report from a British Working Party of Performance Monitoring in the Public Services has, as one of the recommendations, asked that reported performance measures should always include measures of uncertainty (Bird et al. 2005). A 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, 1999, 2000). Bootstrapping provides bias correction of the efficiency scores and confidence intervals, thus signalling the quality of the estimates of productivity changes.[1] The statistical technique of bootstrapping can also be applied to the Malmquist productivity index estimates for individual offices (Simar and Wilson 1999).

The literature on efficiency and productivity of tax offices is rather sparse. In a review paper of the public sector by Simpson (2009) there is only one reference to a study of tax offices (Førsund et al. 2006). It may therefore be of interest to briefly review the few papers found. These papers, mostly published in rather low-profile journals, will be reviewed briefly, focussing on the variables and the methods used, and some weaknesses will be pointed out.

González and Miles (2000), focussing on tax evasion, estimate efficiency scores using DEA on a cross-section for a year (1995) for 15 regional tax offices in Spain. It may be the first paper on tax offices using bootstrapping techniques. The variables are one input variable, share of tax inspectors of total employment in a tax office, and two output variables, number of actions taken concerning the tax returns per return, and the debt—regional product ratio per number of actions.[2] The confidence intervals estimated imply that a hypothesis of equal efficiency scores cannot be rejected. A problem of the study (apart from the relevance of the output variables) is that some bias-corrected efficiency scores and some upper limits for the input-orientated efficiency scores are greater than 1, and two confidence intervals for units being 100 % efficient in the initial DEA run have collapsing intervals with 1 as both the lower and upper limit.

In Thirtle et al. (2000) the efficiency perspective is focussed on the efficiency in raising tax revenue at the regional (state) level and the issue of size of jurisdictions from a tax-efficiency perspective. The data are covering years from 1980 to 1992 for 15 states of India. The four input variables are: collection expenditures as a share of State Domestic Product (SDP), SDP at constant prices, agriculture's share of SDP (entered as the inverse to function as an input), and a poverty index. The single output is tax revenue. Using this output yields a special definition of tax efficiency, and cannot serve as an efficiency index for how the resources consumed by an office relates to the core service production of a tax office; i.e. auditing tax returns. The average Malmquist index and its decompositions are shown using regressions for yearly changes to produce average growth rates.

Moesen and Persoons (2002) study efficiency of 289 regional tax offices responsible for personal income tax in Belgium using cross-section data for 1991. Both DEA and free disposal hull (FDH) methods are used. The single input is personnel in full-time equivalents. The four outputs are number of audited returns in category A (wage earners), number of audited returns leading to increased tax in category A, number of audited returns category B (independent professionals), and number of audited returns leading to increased tax in category B. Results both for FDH and DEA with constant and variable returns are presented in figures.

In three partly overlapping papers the same data base for 41 tax offices located in the Lisbon region covering the period 1999–2002 is used to estimate, firstly efficiency using a stochastic cost frontier in Barros (2005), secondly Malmquist indices of productivity in Barros (2006) and finally estimating efficiency using a cost frontier estimated using DEA in Barros (2007).

In Barros (2005) a stochastic frontier cost function with composed error specified as a Cobb–Douglas functional form in two outputs and three inputs is estimated for the panel data. The outputs are total tax collected (constant 1999 prices) and total clear-up rates of disputes. The inputs are price of labour; total salary of number of equivalent employees, price of capital proxied by rents per unit office space and price of capital proxied by total personal taxes per population in the jurisdiction of the tax office. The cost variable is total operational costs in constant 1999 prices. There is a formal problem identifying scale properties because the estimated cost function is homogenous of degree 2.9 and it is elementary micro knowledge that an optimised cost function is homogenous of degree 1 in prices.

[1] However, this bootstrapping only overcomes the inherent sampling bias of data sets and do not deal with measurement error or model misspecifications. The stochastic frontier approach based on a parametric frontier yields in this respect a more complete statistical analysis. However, the notion of a production function in a technical sense is rather vague for tax offices. Imposing a specific functional form may then be too restrictive and the non-parametric DEA approach is felt to be the preferred approach under such circumstances.

[2] The definition of this variable is not clear.

In Barros (2006) the number of variables is greatly increased to 11 output variables. Tax revenues are disaggregated into six categories and clear-up rates into five categories. The three input variables are number of employees, capital measured by rent and the tax population. All monetary values are in constant 2000 prices. A Malmquist productivity index and its decomposition terms are estimated. The Färe et al. (1992) specification is used, with decomposition into efficiency change and frontier change, but the scale assumption used is not discussed. To believe in technical change over so few years for a service operation characterised by very generic capital structures must be defended better than done in the paper. Managerial skills are not commonly classified as technology as done in the paper, but as an explanation of efficiency.

In Barros (2007) a frontier cost function is estimated using DEA on the same set of variables as in Barros (2006). A price per inhabitant is introduced in the cost function measured as personal taxes per inhabitant of the tax district. This construction seems rather odd. However, the DEA cost model is not explicitly set up, so we are left somewhat in the dark, and no results for the cost efficiency scores for the four years are presented. The consequences of operating with 14 dimensions show up with *all* units being on the frontier in the case of variable returns to scale (VRS).

In Katharaki and Tsakas (2010) the focus is on the total tax that tax offices collect in Greece. The objective is to estimate technical efficiency and scale efficiency using a DEA model (both Charnes et al. 1978 and Banker et al. 1984 models) for 27 tax offices over the period 2001–2006 in the sample selected from mainly agricultural regions. The two output variables are tax from 'natural' persons and tax from legal entities. The four input variables are labour in each office measured by number of employees, number of computers operating in each tax office, number of natural persons paying taxes, and number of legal entities paying taxes. A rather ad hoc window analysis is conducted instead of trying to do a productivity analysis proper.

The papers by Moesen and Persoons (2002), Barros (2006, 2007), and Katharaki and Tsakas (2010) all use a two-step procedure specifying a Tobit censored model regressing efficiency scores from the first step on additional explanatory variables. Moesen and Persoons use as explanatory variables dummy variables for management qualifications and new directorates, number of fines given, number of official assessments, number of control visits, and number of people liable for the personal income tax per tax office. In Barros (2006, 2007) the second step of explaining efficiency are formally identical, and the papers use cost efficiency scores as the dependent variable and selecting variables that are independent of the balance sheet from which the variables used in the DEA analysis are taken. Data on the DEA scores are not presented. One may think that a 14-dimensional specification could create problems. Katharaki and Tsakas use as explanatory variables the population belonging to a tax office relative to the largest population of a tax office, the taxes from services relative to total taxes, the ratio of number of legal entities to the total number of tax entities, and the tax revenue per tax payer and the average gross income in the tax region.

The crucial problem of whether explanatory variables are only influencing efficiency scores but not the production possibility set is not discussed in the first three papers and only mentioned in the last without any argumentation about what kind of relationships there are between efficiency or production possibility set and explanatory variables. A two-stage procedure is only without bias if the explanatory variables are exclusively influencing the efficiency score, but not the production possibility set.

The point of departure for the present study is Førsund et al. (2006) using data for the period 2002–2004 provided by the Norwegian Directorate of Taxes. The key output variables used and the single input variable (total costs) were compiled and selected by the Directorate. The emphasis was on efficiency and confidence intervals for the efficiency scores, but also the productivity development between 2002 and 2004 was briefly studied, using a Malmquist productivity index.

The present study focusses exclusively on productivity change and adds as further developments new methods for visualizing productivity change, including the confidence intervals and a classification into units with significant productivity decline, insignificant change, and significant increase in productivity. A visualization of type of productivity change and change in the single input taking care of the size distribution of units is also developed. These visualisations should have a general interest also for other sector applications.

Main findings are that the productivity change distribution for the total period 2002–2004 ranges from a 20 % decline to a 35 % increase using the point estimates of the bias-corrected productivity changes. Taking the productivity results at face value indicate that both units having a productivity decline and units having a productivity improvement represent about 50 % of the cost in 2004 over the three years. Key information about uncertainty is provided by testing whether change is significantly positive or negative. It then turns out that about 18 % of cost is used in units having a significantly declining productivity, 44 % insignificant change and 38 % in units having significantly increasing productivity. This kind of more nuanced information than point estimates should be useful for policy makers.

The paper is organised in the following way: Sect. 2 presents the methods used for estimating the productivity changes, including briefly bootstrapping. In Sect. 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, illustrated using visualizations, are discussed in Sect. 4 and Sect. 5 concludes.

2 Methodology

2.1 Productivity

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 developed for discrete time based on the ratio of distance functions measures for two observations (e.g., the same unit measured for two different time periods). The frontier function is the same for the two observations. It is well known that distance functions correspond to Farrell (1957) efficiency measures. We will use the latter terminology below.

When applying the Malmquist productivity index attention should be paid to desirable properties (Färe et al. 2008). In the literature this is more often than not glossed over. We will therefore explain in more detail the choice of our specification. Productivity as measured by the Malmquist index 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 should not change productivity. Therefore the benchmark envelopment of data if we want to measure total factor productivity (TFP) is one that is homogenous of degree 1 in the input–output vector, and thus the linear-homogenous set that fits closest to the technology. The homogenous envelopment can be used to define the concept of technically optimal scale (Frisch 1965), termed TOPS in Førsund and Hjalmarsson (2004a,b). This is the scale where the elasticity of scale is 1, and is illustrated in Fig. 1. The constant returns to scale (CRS) line is tangent to the VRS frontier at the point P_2^{tops}.[3] A proper TFP measure is thus obtained by only using the information about changes in the technically optimal scale over time. From classical production theory we know that the productivity is maximal at optimal scale where returns to scale is one, thus this is a natural reference for productivity changes over time.

An illustration in the two-period case is provided in Fig. 1. Observations of the same unit for the two periods 1 and 2 are indicated by P_1 and P_2. The two corresponding VRS frontiers are drawn showing an outward shift indicating technological progress. The TOPS point for period 2

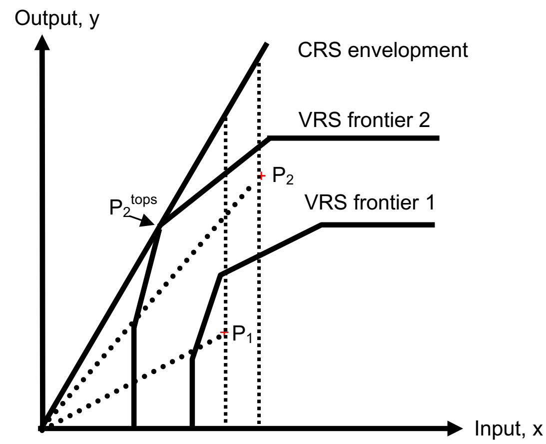


Fig. 1 The Malmquist productivity index. Productivity change for a unit measured relative to optimal scale of the benchmark technology

is labelled P_2^{tops}. Just as the productivity should be unchanged if the input–output vector is proportionally scaled, a measure of productivity should double if outputs are doubled and inputs are kept constant, and should correspondingly halve if inputs double, but outputs are constant. The desirable homogeneity properties of a TFP index is therefore to be homogenous of degree 1 in outputs in the second period and of degree (-1) in inputs of the second period, and homogenous of degree (-1) in outputs of the first period and homogenous of degree 1 in inputs of the first period. Using CRS to envelope the data is thus one way of obtaining all the required homogeneity properties of a Malmquist productivity index.

Another property of a productivity index that is important is the *circularity* of the index (Berg et al. 1992) [see Gini (1931) for an interesting exposition]. The implied transitivity of the index means that the productivity change between two non-adjacent periods can be found by multiplying all the pairwise productivity changes of adjacent periods between the two periods in question. We will transtivitise the Malmquist index by using a single reference frontier enveloping the pooled data. In Tulkens and van den Eeckaut (1995) this type of frontier was termed the *intertemporal frontier*.[4] As is common with indices, performance is calculated using information that may not have been available in earlier periods, but this is consistent with a retrospective evaluation.[5] Using the same CRS reference

[3] In general the technically optimal scale point may not be unique within a piecewise linear technology, 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).

[4] In Pastor and Lovell (2005), missing out on the reference to Tulkens and van den Eeckaut, it is called the global frontier. The authors also claim as new a measure of technical change for a circular index that was originally pioneered in Berg et al. (1992).

[5] An inconvenience is that if one wants to recalculate productivity change after time has elapsed, providing more data, the results for previously studied units may be influenced by the new frontier benchmark, and thus results for the same time periods may not turn out the same.

frontier for all units means that we have made sure that efficiency for all units and time periods refer to the same frontier. Specifying CRS only is not sufficient to ensure that a specific data point occurring at different time periods get the same efficiency evaluation because both input- and output isoquants may differ in shape over time if the technology is allowed to change over time as in Färe et al. (2008). It does not help to take the geometric mean of time-adjacent frontier distances.

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

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

where superscript s symbolises that all data is used as the technology reference set. The Malmquist productivity estimator is conditional on the estimator, $\hat{\hat{S}}^s$, of a linear homogeneous envelopment set. The efficiency scores $\hat{E}_i^s$ are calculated for period u and v respectively for each unit i.

2.2 Bootstrapping

The bootstrap procedure is well established in the literature (Simar and Wilson 1998, 1999, 2000). We are following this procedure and will therefore not go into details here. Testing period frontier functions (Banker 1993) VRS was preferred. Choosing the Farrell output-oriented efficiency measure E_2, distributed on (0,1], our resampling (Efron 1979) creating pseudo replicate data sets, is done on the basis of the calculation of efficiency scores relative to the VRS frontier for each time period

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

where E_{2t}^{KDE} is a draw of the kernel density distribution estimated for the efficiency score. This distribution 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.

A new DEA frontier is then estimated on these pseudo observations (x_i, y_i^{ps}). We make 2000 such draws and establish 2000 new DEA frontiers. Now, going back to each run for a pair of periods, the Malmquist productivity index, given by (1), is calculated using the linear homogeneous technology created for the pooled set of all pseudo observations as the benchmark.

Assuming estimators to be consistent, Simar and Wilson (1999) show that the bias can be estimated based on the relationship

$$\begin{gathered}\left(\tilde{M}^s(u,v) - \hat{M}^s(u,v)\right)|\hat{S}^s \sim \left(\hat{M}^s(u,v) - M^s(u,v)\right)|S^s,\\ u,v = 1,..,T, u \neq v\end{gathered} \tag{3}$$

Here M^s is the true unknown productivity, $\hat{M}^s$ 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. The confidence intervals are based on the estimates of the biases (Simar and Wilson 1999).

3 Data

The local tax offices in Norway, 98 in all, use about 60 % of all labour of the Tax Administration, and are responsible for tax assessment for all types of income tax. (Collection of taxes is done by other organisations). 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 tax authority. 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 using pre-existing data. In view of the 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 Norwegian Directorate of Taxes 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 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 are not based on input costs, but constitute independent quantity measurements and thus may be used for such studies (Førsund et al. 2006).

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 operating cost x net of cost due to special circumstance of the offices, such as extra

Table 1 The data

Variable	Year	Minimum	Maximum	Sum	Mean	SD
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 cost	2002	2,804,888	171,593,294	1,226,341,592	12,513,690	18,291,279
	2003	2,998,829	177,198,456	1,250,478,743	12,759,987	18,985,940
	2004	2,884,602	172,265,680	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	602,963	6153	10,686
	2003	633	101,186	611,812	6243	11,127
	2004	804	110,497	643,080	6562	11,973
Y2: number of false registrations detected by control activities	2002	0	799	3783	39	97
	2003	0	1526	4701	48	156
	2004	0	3299	6925	70	337
Y3: number of tax returns from employees and pensioners	2002	5361	418,785	3,384,913	34,540	46,818
	2003	5604	422,115	3,452,177	35,226	47,531
	2004	5601	428,822	3,462,748	35,334	48,015
Y4: number of complaints on tax assessment	2002	40	16,295	63,407	647	1839
	2003	9	10,018	52,573	537	1211
	2004	9	11,178	48,680	497	1245
Y5: number of returns from non-incorporated businesses	2002	801	32,510	316,542	3230	3411
	2003	824	33,695	325,165	3318	3522
	2004	791	34,722	323,610	3302	3669
Y6: number of corporate tax returns	2002	226	33,264	159,189	1624	3484
	2003	231	31,253	159,908	1632	3304
	2004	267	31,461	162,164	1655	3338

need for transport and rent.[6,7] Six outputs y are specified representing the main activity areas that are to process 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 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.

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

[7] Labour is the dominating input, counting for about 60–80 % of total costs, as is also the case for Belgian tax offices reported in Moesen and Persoons (2002).

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 (1) 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 Sect. 2 (Simar and Wilson 1999), but without assuming correlation between the Malmquist indices over time.

4.1 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

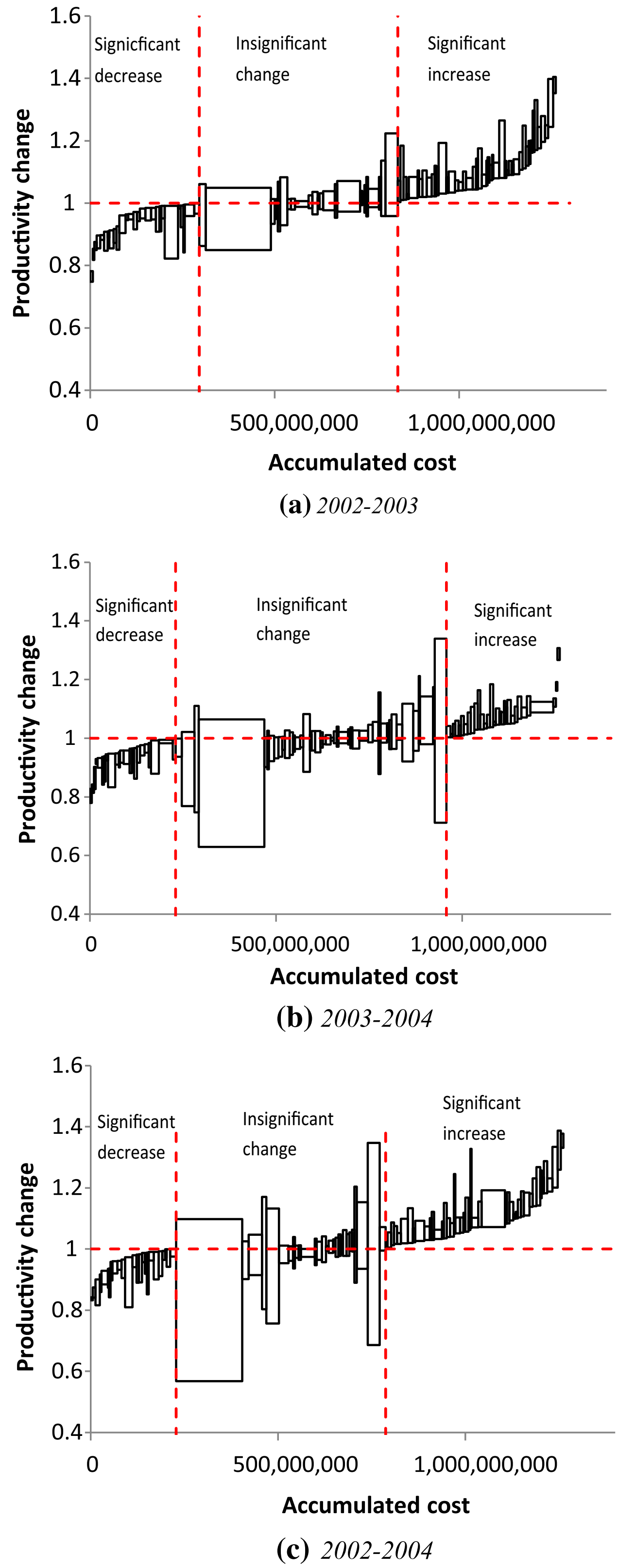
Table 2 Average productivity growth rates in percentage

Period	Growth measure	Original point estimate	Bias-corrected	95 % confidence interval
2002–2003	Average unit	0.8	0.8	0.5–2.0
	Mean	3.7	3.6	3.5–4.2
2003–2004	Average unit	2.1	0.8	−1.7 to 2.6
	Mean	0.3	0.5	−0.4 to 0.7
2002–2004	Average unit	3.0	1.6	0.5–4.3
	Mean	4.3	3.9	3.6–4.7

the calculations (Førsund and Hjalmarsson 1979). Another more conventional approach is to take some mean of the individual results. (Li and Cheng 2007). We have done both approaches and the results are set out in Table 2. 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 quite different for the two types of estimates of productivity, showing a 0.8 % growth for the average unit-measure and 3.7 % for the mean value of the individual estimates. The confidence intervals for both measures show significant growth. Both bias-corrected measures are quite close to the uncorrected point estimates. For the second period, however, the growth figures for the measures are reversed with an average unit measure of 2.1 % and a mean measure of 0.3 %. The biased-corrected average unit measure is rather low at 0.8 % and closer to the mean measure of 0.5 %. But both confidence intervals contain the value of zero implying that the estimates are not significant. From the first to the last period the average unit measure is smaller than for the mean measure and the difference between the values of unadjusted and bias-corrected measures for the former is larger than for the latter measure. The use of an average unit is founded on calculating how well the sector in total keeps up with the best of its own constituent units (Farrell 1957), but the mean measure may be most suitable for reflecting the average development.

4.2 Productivity development for individual units

Testing hypothesis whether an office has had a significant decline or increase in productivity, as stated in Simar and Wilson (1999), is one of the benefit of bootstrapping. We have in Fig. 2, in order to show the productivity results and illustrate the testing at the same time, set out new *productivity significance diagrams* focussing on the confidence



(a) *2002-2003*

(b) *2003-2004*

(c) *2002-2004*

Fig. 2 Significance testing: units grouped by the nature of the significance of productivity change. Sorted by lower limit, mid-point, and upper limit of confidence intervals respectively. Width of boxes proportional to cost in 2002

Table 3 Productivity change (bias corrected) and cost shares 2004. Significance level 95 %

Periods	Productivity decrease				Insignificant change		Productivity increase			
	Point estimate		Significant				Point estimate		Significant	
	Cost share	# units	Cost share	# units	Cost share	# units	Cost share	# units	Cost share	# units
2002–2003	50	36	23	27	43	25	50	62	34	46
2003–2004	52	46	20	24	56	43	48	56	24	31
2002–2004	50	39	18	25	44	28	50	59	38	45

intervals for the units. The width of a box is based on the relative share of costs for 2002. The height of the box shows the width of the confidence interval. The bias-corrected productivity value is not shown explicitly, but may be visualised as the mean of the confidence intervals.

The units are grouped in three groups, starting from the left with units with significant decrease in productivity, then units with insignificant productivity change, and lastly 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,[8] 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 for the second group illustrates the position of the interval relative to the crucial value of 1 signifying no productivity change.

In Fig. 2a we have that in the group having significant productivity decrease there are 25 out of 98 units representing 38 % of costs. The cost shares and the number of units using point estimates (bias corrected) of productivities together with using the information from confidence intervals, enabling significance evaluations, are set out in Table 3. There is a relative over-representation of small units in the group with significant decrease. The group of insignificant change of 25 units represents 43 % of costs and has three of the largest units and also some medium-sized units. The large and medium-sized units in this group tend to have the widest confidence intervals. The last unit in the right-hand tail has the largest confidence interval of all units, and have a point estimate of 1.16 of the Malmquist productivity index. The group of significant increase have the largest number of 46 units representing only 34 % of total cost, pointing to an overrepresentation of small units.

Figure 2b reveals a structural change from the period 2002–2003 to the period 2003–2004. The group with significant decrease has decreased in number from 27 to 24. The share of costs has decreased from 23 to 20 %, implying a stable average size of units in this group. Both the number of units and the share of costs have increased markedly in the group of insignificant change; the number from 25 to 43 and the share from 43 to 56 %. The average size of units in the group has thus increased considerably. Both the share of costs and the numbers have been reduced markedly in the group with significant increase in productivity; from 34 to 24 %, and from 46 to 31 for shares and units, respectively, indicating a smaller average size. The shrinking of both the groups with significant decline and significant increase, and the increase in the group of insignificant change, reflects the general tendency of a widening of the confidence intervals.

The strength of the Malmquist productivity index approach is that it gives us detailed information about the development of productivity for each observation. Having a common reference frontier for all time periods allows us to study the development of units using Fig. 2. Comparing Fig. 2a and b we see that the largest unit remains in the group with insignificant change, but that the point estimate of productivity change has been reduced due to the width of the confidence interval increasing substantially for medium sized and large units. The second-largest office has moved from the group of insignificant change to the group of significant increase, thus having a genuine increase in productivity. Its confidence interval has become quite narrow. The large unit at the right-hand end of the distribution of insignificant units is in the same position, but now with a wider confidence interval, in fact the widest of all units. There are in general quite a number of movements of the productivity results of units, pointing to some turbulence as to the match between exogenously given tasks to perform and the manpower allocated to the offices. This is especially the case for the small units.

A way of exhibiting the "micro foundation" of the average development seen in Table 2 is to show the overall 2002–2004 development as seen in Fig. 2c. Having only three periods this may not be so interesting, but having more observations such a diagram from the first period to

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

the last may give us a quick feel for the general development. A characteristic of the first–last period picture is that many of the confidence intervals have become wider, especially for large units in the group with insignificant change, and the numbers of units in each group reflect more an average situation as revealed in Table 3.

Lower 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 smaller units being more numerous, but that we must be careful when using productivity figures for the large units. This is especially evident for the largest unit in Fig. 2c. 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 from 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 characterisation of the nature of productivity growth (Førsund and Kalhagen 1999; Førsund et al. 2006). In Fig. 3 productivity change for 2002–2003, 2003–2004 and 2002–2004 is shown together with the relative change in costs. The area of a circle is proportional to costs in 2002. The open circles are the units with significant productivity change (either negative or positive), while the circles with grey fill are units with insignificant change, corresponding to the situation revealed in Fig. 2. The horizontal axis measures cost change, and the vertical axis measuring productivity change is placed at zero change in costs. To the left of the origin the costs have decreased while to the right costs have increased. The total range for the two years and the first year to the last vary from −24 to +33 %, −24 to 18 %, and −22 to 19 % in Fig. 3a–c, respectively. The horizontal line at the value 1 delimitates the units with productivity decrease and increase, respectively, and the vertical axis from zero change in costs form four quadrants numbered I to IV.

In Fig. 3a units in Quadrant I have had both productivity growth and increase in cost. Such units may be said to have experienced *efficient cost increase*. There are few units in this quadrant with positive point estimate for both dimensions, but only one unit with a significant growth in productivity of 6 % and a cost increase of 1 %. The units in Quadrant II have also had productivity growth, but experienced cost reductions. This may be termed *efficient cost saving*. This quadrant has the highest number of units with most of them with significant increase in productivity. The unit with the highest productivity change has had an increase of 38 % (maximal of all units) and reduced cost with 24 % (also maximal). In quadrant III productivity decrease is combined with labour decrease. This is *inefficient cost*

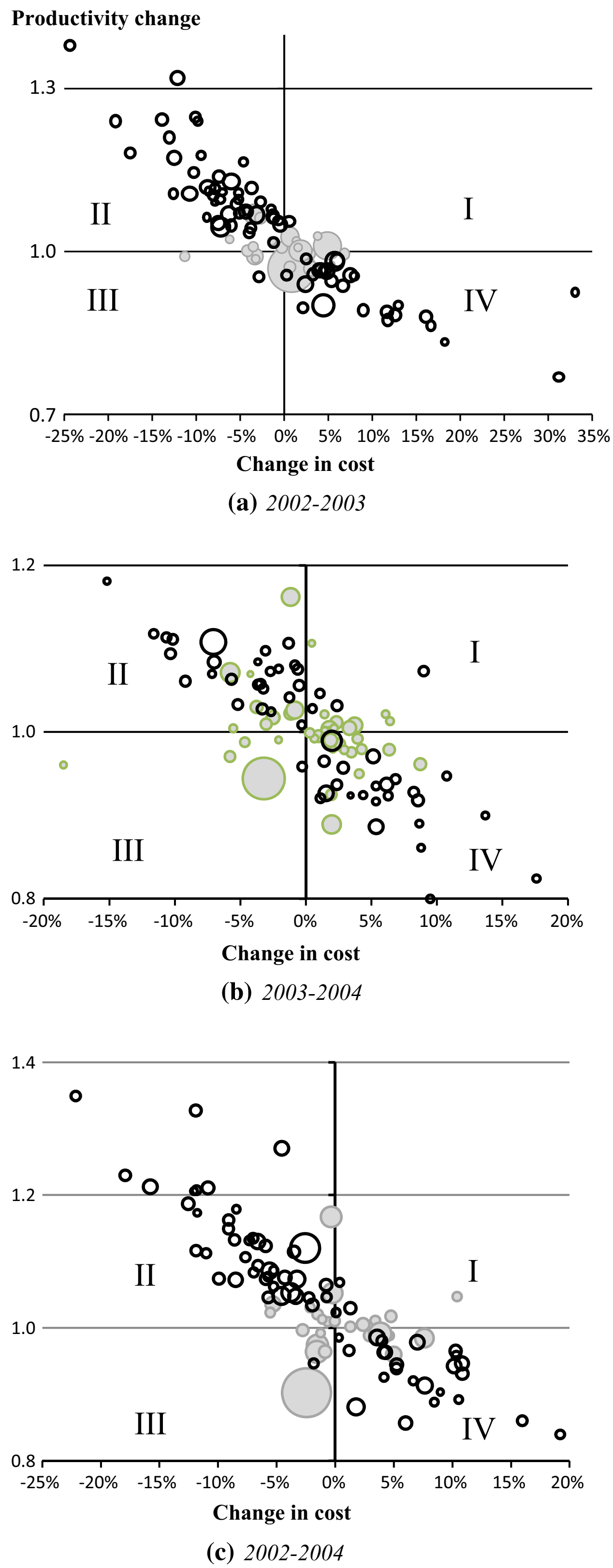


Fig. 3 Productivity and cost change 2002–2004. Size of circles is proportional to cost in 2002

saving. There are relatively few units in this quadrant and again only one unit with a significant decrease in productivity of 4.5 % and a cost decrease of 3 %. Units in Quadrant IV have the worst of both worlds with decreasing productivity and increasing costs. This is *inefficient cost increase*. The majority of units have significant productivity decrease, but here are also units with insignificant change. The unit with the greatest change in costs of 33 % and a decline in productivity of 7.5 %, while the unit with the greatest productivity decline of 8 % has had a cost increase of 31 %.

A few units are extreme in their change in costs, like the two units in Quadrant IV furthest to the right and also the three units in Quadrant II furthest to the left. This may be explained by reorganisation and moving of tasks between offices. However, there is also the possibility of problems with data quality because a small unit in question change both cost and productivity quite drastically going from a cost reduction of 24 % and significant productivity increase of 38 % to a cost increase in the next period of 3 % and an insignificant productivity change of 2 %. Other small units go from cost increases in the range of 33–18 % with significant decrease in productivity to decrease in costs in the range of 18–12 % and significant positive productivity changes in the range of 18–12 %.

Figure 3a revealed that large units have rather small changes in costs and insignificant change in productivity, while for the next period 2003–2004 in Fig. 3b the changes in both costs into negative values and productivities are greater with one unit (second largest) getting a significant increase in productivity of 11 % with a substantial cost decrease of 7 %. The largest unit still has insignificant change in productivity but has now moved from Quadrant IV to Quadrant III according to the point estimate.

Figure 3c looks similar to Fig. 3b, but as pointed out there is a systematic movement of small units from Quadrant II to Quadrant IV. This last point is lost if diagrams from the first period to the last is used as expressing the average change over time.

Because productivity is a ratio between outputs and inputs it is to be expected that there is negative correlation between productivity and costs. This is evident from all panels.

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. For external use a measure the aggregate productivity performance may be of interest, like reporting the figure of 4 % productivity increase in Table 2 on the average over the three-year period. For internal use revealing the productivity performance of individual units, as seen in Fig. 2, should be useful. The results can give valuable information for further investigations trying to explain differences in productivity performance across micro units.

Results of performance measurement of units should be presented in ways that may contribute to promoting 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 it is crucial to use methods that enable us to make a statistical assessment of at least part of the uncertainty of productivity estimates that are the 'engine' of performance measurement over time. The results in Figs. 2 and 3 for the periods 2002–2003 and 2003–2004 show that large units would have appeared to show 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 cost share of units with point estimates of productivity increase declined from 50 to 48 %, and the share with apparent productivity decrease increased from 50 to 52 %, but declined from 23 to 20 % when looking at the significant changes only. The range of productivity changes are from −25 to +38 %, and resulting in an overall productivity increase of about 4 % for 2002–2004 for the mean. The range of change may seem somewhat surprising for such a short period.

The productivity results reveal in general 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. For policy implications it should be noted that the confidence intervals for the large units are wide, while they are narrower for small units. This is mainly due to the less central position in the dataset of the relative few large 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 due to product mix, and warrant clarification by further research.

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 Figs. 2 and 3, 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 RD (1993) Maximum likelihood, consistency and data envelopment analysis: a statistical foundation. Manag Sci 39(10):1265–1273

Banker RD, Charnes A, Cooper WW (1984) Some models for estimating technical and scale inefficiencies. Manag Sci 30:1078–1092

Barros CP (2005) Performance measurement in tax offices with a stochastic frontier model. J Econ Stud 32(6):497–510

Barros CP (2006) Measuring total productivity in Lisbon tax offices with a Malmquist index. Tijdschrift voor Econo en Manag 51(1):25–46

Barros CP (2007) Technical and allocative efficiency of tax offices: a case study. Int J Public Sect Perform Manag 1(1):41–61

Berg SA, Førsund FR, and Jansen ES (1992) Malmquist indices of productivity growth during the deregulation of Norwegian banking, 1980-89. The Scand J Econ 94, Supplement. Proceedings of a symposium on productivity concepts and measurement problems: welfare, quality and productivity in the service industries, S211–S228

Bird SM, Cox Sir D, Farewell VT, Goldstein H, Holt T, Smith PC (2005) Performance indicators: good, bad, and ugly. J R Stat Soc Ser A 168(Part 1):1–27

Caves DW, Christensen LR, Diewert E (1982) The economic theory of index numbers and the measurement of input, output, and productivity. Econometrica 50(6):1393–1414

Charnes A, Cooper WW, Rhodes E (1978) Measuring the efficiency of decision making units. Eur J Oper Res 2(6):429–444

Efron B (1979) Bootstrap methods: another look at the jackknife. Ann Stat 7:1–6

Färe R, Grosskopf S, Lindgren B, Roos P (1992) Productivity changes in Swedish pharmacies 1980–1989: a non-parametric Malmquist approach. J Prod Anal 3:85–101

Färe R, Grosskopf S, Margaritis D (2008) Efficiency and productivity: Malmquist and more. In: Fried HO, Lovell CAK, Schmidt SS (eds) The measurement of productive efficiency and productivity growth, vol 5. Oxford University Press, New York, pp 522–622

Farrell MJ (1957) The measurement of productive efficiency. J R Stat Soc Ser A 120(3):253–281

Førsund FR, Hjalmarsson L (1979) Generalised Farrell measures of efficiency: an application to milk processing in Swedish dairy plants. Econ J 89:294–315

Førsund FR, Hjalmarsson L (2004a) Are all scales optimal in DEA? Theory and empirical evidence. J Prod Anal 21(1):25–48

Førsund FR, Hjalmarsson L (2004b) Calculating scale elasticity in DEA models. J Oper Res Soc 55:1023–1038

Førsund FR, Kalhagen KO (1999) Efficiency and productivity of Norwegian colleges. In: Westermann G (ed) Data envelopment analysis in the service sector. Deutscher Universitäts-Verlag, Wiesbaden, pp 269–308

Førsund FR, Kittelsen SAC, Lindseth F, Edvardsen DF (2006) The tax man cometh-but is he efficient? Natl Inst Econ Rev 197(July):106–119

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

Gini C (1931) On the circular test of index numbers. Metron 9(2):3–24

González MX, Miles D (2000) Eficiencia en la inspeccion de hacienda. Revista de Economia Aplicada 8(24):203–219

Grifell-Tatjé E, Lovell CAK (1995) A note on the Malmquist productivity index. Econ Lett 47:169–175

Katharaki M, Tsakas M (2010) Assessing the efficiency and managing the performance of Greek tax offices. J Adv Manag Res 7(1): 58–75

Li S-K, Cheng Y-S (2007) Solving the puzzles of structural efficiency. Eur J Oper Res 180:713–722

Malmquist S (1953) Index numbers and indifference surfaces. Trabajos de Estadistica 4:209–224

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

Pastor JT, Lovell CAK (2005) A global Malmquist productivity index. Econ Lett 88:266–271

Silverman BW (1986) Density estimation for statistics and data analysis. Chapman and Hall, London

Simar L, Wilson PW (1998) Sensitivity analysis of efficiency scores: how to bootstrap in nonparametric frontier models. Manage Sci 44:49–61

Simar L, Wilson PW (1999) Estimating and bootstrapping Malmquist indices. Eur J Oper Res 115(3):459–471

Simar L, Wilson PW (2000) Statistical inference in nonparametric frontier models: the state of the art. J Prod Anal 13:49–78

Simpson H (2009) Productivity in public services. J Econ Surv 23(2): 250–276

Thirtle C, Shankar B, Chitkara P, Chatterjee S, Mohanty MS (2000) Size does matter: technical and scale efficiency in Indian state tax jurisdictions. Rev Dev Econ 4(3):340–352

Tulkens H, van den Eeckaut P (1995) Non-parametric efficiency, progress, and regress measures for panel data: methodological aspects. Eur J Oper Res 80:474–499