A Frontier-based System of Incentives for Units in Organisations with Varying Degrees of Decentralisation

Mohsen Afsharian
Corresponding author. Department of Business Sciences, Technische Universität Braunschweig, Fallersleber-Tor-Wall 23, 38100 Braunschweig, Germany, m.afsharian@tu-bs.de

Heinz Ahn
Department of Business Sciences, Technische Universität Braunschweig, Fallersleber-Tor-Wall 23, 38100 Braunschweig, Germany, hw.ahn@tu-bs.de

Emmanuel Thanassoulis
Department of Operations and Information Management, Aston Business School, Aston University, Birmingham B4 7ET, United Kingdom, e.thanassoulis@aston.ac.uk

DISCLAIMER
THIS PAPER WAS ACCEPTED (2018) FOR PUBLICATION IN THE EUROPEAN JOURNAL OF OPERATIONAL RESEARCH. CERTAIN DETAILS MAY DIFFER BETWEEN THIS AND THE PUBLISHED VERSION .OF THE PAPER

Abstract.

The paper focuses on hierarchically structured organisations with a large set of operating units. While the central body in such organisations faces asymmetry of information concerning the operating costs of the units, it may wish to incentivise them through benchmarking and target setting to operate as efficiently as possible. If a standard Data Envelopment Analysis (DEA) approach is used for this purpose, each operating unit could estimate its own efficient targets. However, this decentralised scenario is not necessarily appropriate for a centralised organisation in which a central body wishes to optimise the performance of the system of units as a whole. On the other hand, a top-down imposed set of targets is often not suitable as they would be too demanding for some units and too lax for others. This paper proposes a DEA-based approach for incentivising the units of a hierarchically structured organisation in order to optimise the performance of the units collectively while at the same time the targets are not too demanding for inefficient units. The proposed approach is also extended so that incentive levels for operating units are determined over time, taking into account any changes in their productivity. Accordingly, the central management can strike a balance between not spending too much on incentives on the one hand and encouraging the operating units to reveal their true cost function on the other. We illustrate our approach using data from a set of German savings banks.

Keywords: Data Envelopment Analysis (DEA); Incentive Regulation; Centralised Management; Clustering; Banking

1. Introduction

In many instances in the public and the private sector, we encounter situations where a central body manages a large set of similar production units. Examples of these centrally managed multi-unit organisations are a bank managing its branches, a tax authority managing local tax offices, a supermarket chain managing its outlets and so on.[1] In such organisations, the central management differs substantially between organisations in the degree to which it controls the day-to-day business of the operating units and imposes instructions on them. Operating units under different degrees of decentralisation are given different levels of autonomy in terms of how, e.g. they can take actions, make decisions and cooperate with others to deliver services to customers. The local managers may accordingly focus on different principles for decision-making such as individual goals and strategies which might not be optimal for the organisation as a whole. Therefore, the need arises for the central management to develop an appropriate incentives system which reflects its specific degree of decentralisation and can encourage the local management of each unit to act in a way which optimizes performance for the organisation as a whole. The following example from the German banking system can illustrate more precisely the organisational structure outlined above.

German savings banks, with the common brand name *Sparkasse*, are essentially credit institutions whose activities focus on providing financial services for private individual customers as well as for small and medium-sized enterprises within their specified geographic areas (see Vitols 1995; Simpson 2013). These banks are managed in a self-reliant way and locally administrated by their own management board which is responsible for the day-to-day conduct of their business. Nevertheless, the savings banks are also controlled centrally by the German Savings Banks Association (*Deutscher Sparkassen- und Giroverband, DSGV*). As an umbrella organisation, DSGV strives to encourage inefficient banks to become efficient and those with good efficiency are incentivised to continue being so (see Afsharian and Ahn 2016; dsgv.de 2017b). Towards this end, a transparency-based incentive method is being applied by DSGV. Different financial- and credit-based ratings (e.g. those from rating agencies such as DBRS, Fitch and Moody's) are continuously monitored and reported annually to the savings banks and their stakeholders (dsgv.de 2017a). The challenge for DSGV would, however, be to run a more explicit incentives system coupled with an appropriate efficiency measurement mechanism which incentivises the savings banks to a better performance by controlling, e.g., their centrally allocated operating budgets or any other funds available in the group.

[1] In a broader view, an appropriate modification of the approach being developed in this paper can be used in cases in which there exist natural monopolies instead of usual competitive markets (see Section 2). Examples are those of large infrastructure industries like water, electricity and gas networks, e.g., see Thanassoulis (2000) concerning water distribution, Førsund and Kittelsen (1998) concerning electricity distribution, Hawdon (2003) concerning gas distribution as well as Agrell and Bogetoft (2001) concerning healthcare.

On the one hand, DSGV does not have access to full information as to the true cost function that pertains to each bank in delivering products and services locally (such as savings and transactional accounts, personal loans, processing financial documents and providing advisory services). This leads to an asymmetry of information which can be exploited by the savings banks to extract rents, i.e. the banks may slacken effort actually needed to be cost-efficient. On the other hand, the particular organisational structure of the savings banks does not assign all power to decide about allocating resources and offering products or services to the central DSGV management. This liberal management framework induces conflicts of interest as to which actions should be taken by savings banks locally. As a consequence, adverse selection and moral hazard problems can occur where a savings bank may act inappropriately (from the viewpoint of DSGV) if its interests differ from those of the central management. For example, a saving bank located in an industrial area may focus on services to companies although DSGV may have determined the "concentration on individual consumers" as the most promising market strategy for the group.

This paper adds to the literature by introducing a method for incentivising operating units to act in the best interests for their hierarchically structured organisations like the group of German savings banks outlined above. In the course of a particular degree of decentralisation, the units are operating under a supervising management, but there is a significant pre-given level of rights and flexibilities concerning their local activities, i.e. the organisational structure is neither fully decentralised nor fully centralised.

The proposed method draws among others from the frontier-based incentive mechanism in the context of regulation suggested by Bogetoft (1997), which has its roots in the seminal work of Shleifer (1985). In order to measure efficiency under a pre-specified degree of decentralisation, an adaptation of the approach introduced by Thanassoulis (1996) is first applied to cluster operating units by the ratios of their output levels ("output mix"). A set of common weights (e.g. cost levels per unit of output) is then determined allowing to incentivise the operating units in a manner which most closely reflects their own output profiles and cost structures. This method also provides a platform by which the central management can decide about the level of cost savings desired from the inefficient units and the rewards given to units with a good efficiency. Finally, the proposed method is extended by a multi-period approach. This is done by means of a new framework of the Malmquist index inducing also the inefficient units to save costs to keep up with the moving efficient boundary which reflects productivity gains expected in the future.

The paper proceeds as follows: Section 2 gives a brief overview of the use of data envelopment analysis (DEA) for incentivising operating units. In Section 3, we present a new approach – in both its static and multi-period version – for incentivising the units of a centrally managed multi-unit organisation under different degrees of decentralisation. Section 4 illustrates our approach using data from the group of German savings banks. Section 5 concludes the paper.

2. Motivations and Preliminaries

Let us assume that we have a centrally managed multi-unit organisation with a particular degree of decentralisation, like the one of German savings banks outlined in the introduction. In this context, the central management or "regulator" (e.g. DSGV) oversees *n* decision making units (DMUs) or "agents" (e.g. savings banks) who may benefit from a natural monopoly or pre-given rights and flexibilities in producing certain products and/or services. Let $Y_j = (y_{1j}, y_{2j}, ..., y_{sj}) \in \Re^s_+$ be a non-zero vector which quantifies the level of outputs of DMU$_j$ (*j*=1,…,*n*). The regulator seeks to avoid the misuse of asymmetry of information or monopoly power where it exists, and to generally incentivise the agents to increase their performance by controlling their budget c_j (*j*=1,…,*n*).

We focus at this stage on the single time period context. It is assumed that there is asymmetry of information between the local autonomous units and the regulator concerning the "technology" by which the budget is spent and converted to outputs, allowing for adverse selection and moral hazard issues. In DEA, this technology is characterized by a few basic context-dependent assumptions and represented by a production possibility set (PPS). Supposing that the regulator does not have a priori information about the details of the cost structure, Bogetoft (1997) proposed the following DEA-based incentive formula by which an optimal compensation plan for DMU$_p$ – indicated by c^*_p – is obtained:

$$c^*_p = c_p\left[1 + \rho_p(\theta_p - 1)\right], \qquad p = 1,...,n. \tag{1}$$

In this formula, c_p represents the observed historical costs of the operations of DMU$_p$. θ_p quantifies the efficiency of DMU$_p$ which is obtained by an appropriate DEA model.[2] Thus, θ_p is the fraction of costs c_p that the activities of DMU$_p$ would actually justify if it had been operating as efficiently as benchmark units. Therefore, $(1-\theta_p)$ is the fraction of c_p available for saving. The parameter ρ_p is user-specified, usually as a fraction of 1. As Agrell et al. (2005) note, ρ_p represents "the power of the incentive scheme" which moderates the savings fraction $(1-\theta_p)$ imposed on DMU$_p$. A good theoretical foundation of the above DEA-based incentive regulation can be found in Bogetoft (1994), Bogetoft (1997) and Agrell et al. (2005).

[2] Depending on the context, one may also use other benchmarking tools to measure efficiency. Examples of such methods which have been applied in incentive regulation are Stochastic Frontier Analysis (SFA) and Stochastic Nonparametric Envelopment of Data (StoNED). See Bogetoft (2013) and Kuosmanen et al. (2015) for an updated overview of SFA and StoNED, respectively.

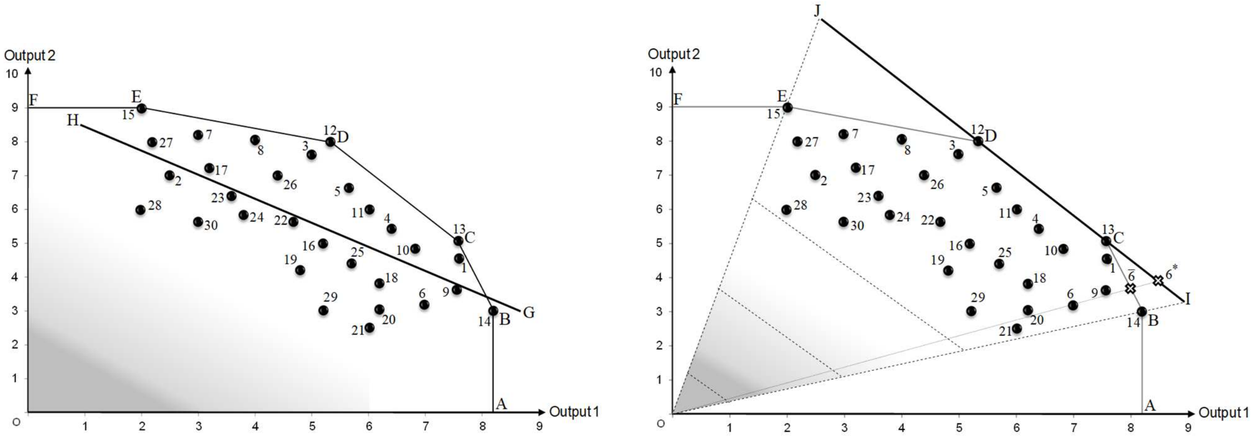
A two-dimensional example helps to illustrate how the above incentives method works under different degrees of decentralisation. Suppose that there exist 30 DMUs (e.g. German savings banks) in the system with two outputs (e.g. total loan and other earning assets) and a single input (e.g. total expenses). Having normalized the data of these units, Fig. 1 represents the output levels of each unit per standard unit of expenditures (costs). Without loss of generality, we postulate that ρ_p=1 in formula (1) and the PPS is characterised by non-emptiness, free disposability, convexity and constant returns to scale (CRS), which can be represented mathematically as

$$PPS = \left\{ (c,Y) \in \Re_+ \times \Re_+^s \,\middle|\, c \geq \sum_{j=1}^{n} \lambda_j c_j, \quad Y \leq \sum_{j=1}^{n} \lambda_j y_{rj}, \quad \lambda_j \geq 0, \quad j = 1,\ldots,n \right\}. \qquad (2)$$

Let us assume that applying (2) on the data set in our example leads to the PPS as shown in Fig. 1(a). This PPS, which is bounded by ABCDEF, consists of all output levels feasible, in principle, for a standard unit of costs. Note that we can express any line (e.g. GH), in Fig. 1(a) with, e.g., an intercept *b* on the y_2 axis, as $\hat{u}_1 y_1 + \hat{u}_2 y_2 = b$. Dividing across by *b* and setting $\hat{u}_1 / b = u_1$ and $\hat{u}_2 / b = u_2$, we have $u_1 y_1 + u_2 y_2 = 1$. It is well known that in the multiplier form of a standard DEA model (like the CCR model proposed by Charnes et al. 1978) for any DMU that is efficient, the sum of weighted inputs equals the sum of weighted outputs (e.g. see Thanassoulis 2001, Chapter 4, about the value-based model 4.5). Thus, in the context of Fig. 1(a) for any DMU such as E or D on the efficient frontier (see ED in Fig. 1(a)), we would have $\tilde{u}_1 y_1 + \tilde{u}_2 y_2 = \tilde{v} \cdot 1$ where 1 is the unit cost level we assumed for all units in Fig. 1. Dividing across by $\tilde{v}$ and setting $\tilde{u}_1 / \tilde{v} = u_1$ and $\tilde{u}_2 / \tilde{v} = u_2$, we arrive at $u_1 y_1 + u_2 y_2 = 1$. In this context, u_1 and u_2 are costs per unit output y_1 and y_2, respectively, so that their aggregate costs are 1. Thus, one way to see what the DEA model does is to identify efficient reference hyperplanes (e.g. BC, CD, ED in Fig. 1(a)) each one being associated with a different set of "efficient" unit cost levels for the outputs (a fuller discussion can be found, e.g., in Thanassoulis 1996).

We shall refer to any line in Fig. 1 which can be written as $u_1 y_1 + u_2 y_2 = 1$ with its own unique costs per unit output u_1 and u_2 as an "incentive map". This is to convey the notion that the unique set of costs per unit output associated witch each line can be used to incentivise operating units for improved efficiency. The method we develop in this paper revolves around the notion that DEA can be used to enable central management to determine incentive maps in a controlled manner to incentivise the operating units to reveal information on efficient practices mitigating in this way the detrimental effects of the asymmetry of information between central management and the operating units.

Figure 1. Representation of incentive maps under the conventional decentralised and centralised perspective



(a) (b)

Let us now illustrate graphically the way that the common – decentralised and centralised – perspectives of measuring efficiency can be used to incentivise units according to (1). Under a decentralised perspective, a standard DEA model like the one of Charnes et al. (1978) yields three incentive maps with relative costs per unit of each output as associated respectively with lines BC, CD and DE in Fig. 1(a). Thus, under the decentralised regime, the efficiency scores and the corresponding incentives are not calculated in a uniform way due to the fact that different units will normally be projected on a different segment of the efficient frontier and so on a different incentive map. This can create a serious perception problem of equity on the part of the units, irrespective of the technical merit of the approach. The problem of perceived inequity would be especially severe in public sector multi-unit organisations (e.g. the funding of schools, public health care provision etc.) where the multitude of outputs would lead to many incentive maps while units may expect a transparent relatively simple assessment structure for being funded by public funds. Moreover, using "boundary benchmarks" as in Fig. 1(a) could lead to the so-called ratchet effect as boundary performance will in subsequent assessments be integrated as "normal" and further efficiencies sought. Thus, boundary units have no incentive to reveal further cost savings that may be feasible, and instead may prefer to enjoy additional costs as "slack" (Bogetoft 1997 and Agrell et al. 2005).

Going to the other extreme of using a single incentive map, where costs per unit output are "efficient" in the sense of inducing efficiency in DEA terms, is also problematic. A common benchmark to measure the efficiency of the full set of operating units has been put forth by a variety of authors (e.g., Roll et al. 1991; Roll and Golany 1993; Lozano and Vila 2004; Kao and Hung 2005; Cook and Zhu 2007; Asmild et al. 2009; Fang 2013; Varmaz et al. 2013; Mar-Molinero et al. 2014; Afsharian et al. 2017). This way of measuring efficiency preserves the consistency of unit output costs across the operating units in the context of incentivisation using formula (1). There is a variety of possibilities for defining a common benchmark. One possibility would be to use only one of the original incentive maps from DEA, like BC, CD or DE in Fig 1. However, this is problematic. For example, using line IJ in Fig. 1(b) as an incentive map, the efficiency scores and the corresponding incentives are calculated in a uniform way, i.e. all units receive a similar set of costs per unit output. However, this line lies partly to the right and above the lines BC and DE. Accordingly, the required cost savings could prove to be too demanding and possibly infeasible as we have no evidence units can operate outside the PPS. As an example, for DMU_6 – keeping its output mix constant – the centralised incentive map CD would require output levels at 6^* on the extended version IJ of CD. Output levels at 6^* are outside the PPS and we have no evidence that it represents a feasible production point. Moreover, in this scenario with a single frontier incentive map, DMUs receive non-positive incentives only and they have no incentives to improve efficiency if they are already on the efficient frontier.

As Agrell et al. (2005) note, an appropriate incentivisation approach should "avoid arbitrariness, excessively high or negative informational rents as well as ratchet effects". Centered on these properties and taking into account the level of autonomy given to the operating units in the system, our approach will provide a platform by which the central management can decide about the desired level of cost savings from inefficient units and the rewards given to units with good efficiency so that they may reveal further efficient levels beyond those identified in a given PPS.

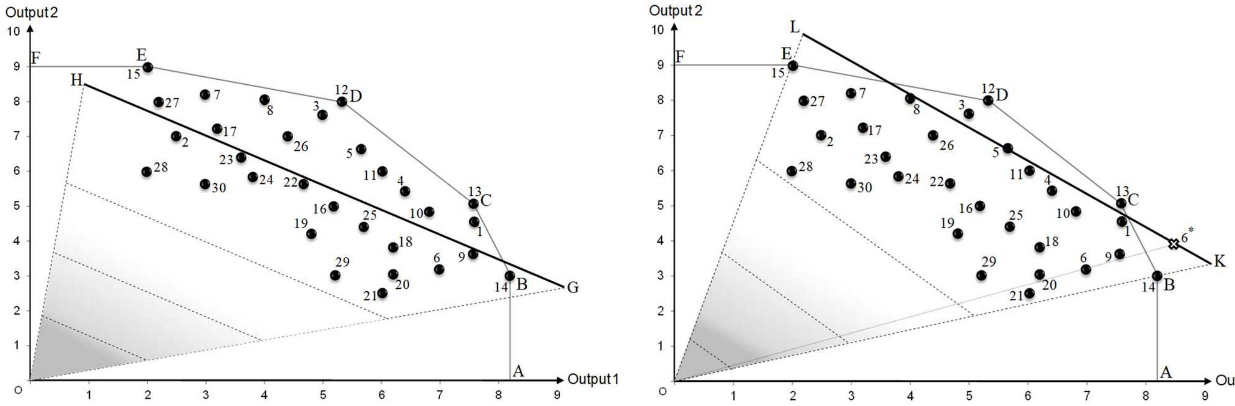
3. The Proposed Incentivisation Method

3.1. The Basic Idea of Determining Incentive Maps

As we saw in the previous section, incentive maps are isocost lines, or more precisely hyperplanes, each one reflecting a different set of costs per unit output. These costs can in turn be used to assess the scope for efficiency savings for some units and potential overcompensation for other units where no efficiency savings are required. The central contribution of the incentivisation method we propose in this paper is to allow for the specification of an appropriate number and location of incentive maps for a given set of DMUs. The larger the number of incentive maps, the higher the degree of decentralisation of a multi-unit organisation. The degree of decentralisation reflects the level of autonomy given to the local operating units in managing resources and outputs, which induces varying cost structures and output profiles within the group of DMUs. Therefore, a framework is proposed by which operating units are clustered, and incentive maps are derived for the units in each cluster. The framework excludes incentive plans which would require a unit to achieve better performance than that determined as feasible by the PPS in (2). Moreover, the incentive plans will be located close to but within the efficient frontier and this in turn will enable well performing DMUs to be overcompensated, incentivising them in this way to reveal further efficiency levels that may be feasible. The basic idea is illustrated by means of the set of DMUs depicted in Fig. 2 which uses the same DMUs as those in Fig. 1.

Let us assume the extreme case that the central management wishes to measure efficiency by grouping its DMUs into a single cluster so that incentives are determined in a uniform way among the DMUs and that all efficient output targets are within the PPS. To operationalize this concept, with respect to the data in Fig. 2(a), one could use the incentive map GH which is obtained as the regression line under the ordinary least squares principle. On this basis, the operating units located below this line will not be fully compensated and thus encouraged to make cost savings. By contrast, the units above GH are overcompensated by being given higher budgets than their projected costs.

Figure 2. Representation of alternative incentive maps under the premise of a single incentive map



(a)

(b)

Although this approach provides an easy to implement way of determining incentives, it suffers from a serious drawback: Due to its central tendency feature, all DMUs' inefficiencies are incorporated in determining the incentive map and thus in deriving incentive plans for the operating units. Fig. 2(a) shows the significant impact of the inefficient DMUs on the determination of the incentive map within this approach. Though one could "shift" the regression line in the sense of setting as benchmarks to say 50% rather than the whole of the excess compensation the regression line yields, the problem remains. The regression line incorporates inefficiencies and it is not clear to what extend it can be shifted in the manner indicated here. (A fuller discussion of drawbacks of such approaches can be found, e.g., in Thanassoulis 1996 and 2001).

In order to overcome this problem, historical inefficiencies should be eliminated from the data to estimate an incentive map. For this purpose, the observed output levels $Y_j = (y_{1j}, y_{2j},..., y_{sj}) \in \Re^s_+$ of DMU$_j$ (j=1,…,n) are replaced by estimated efficient output levels $\tilde{Y}_j = (\tilde{y}_{1j}, \tilde{y}_{2j},..., \tilde{y}_{sj}) \in \Re^s_+$. The latter project the units radially onto the efficient facets of the PPS by $\tilde{y}_{rj} = \sum_{j=1}^{n} \lambda_j^* y_{rj}$, where λ_j^* is the optimal value of λ_j, computed by the following DEA model:

$$\max\left\{\varphi_p+\varepsilon\left(\sum_{r=1}^{s}S_r^+\right)\left|\begin{array}{l}\sum_{j=1}^{n}\lambda_j c_j=c_p\\ \sum_{j=1}^{n}\lambda_j y_{rj}-S_r^+=\varphi_p y_{rp}\qquad r=1,\ldots,s\\ S_r^+\geq 0\quad r=1,\ldots,s;\\ \lambda_j\geq 0\quad j=1,\ldots,n;\ \varphi_p\ \textit{free in sign}\end{array}\right.\right\}. \tag{3}$$

In this model, λ_j (j=1,…,n), S_r^+ (r=1,…,s) and φ_p are variables. ε is the non-Archimedean infinitesimal which is used to prioritise the maximisation of φ_p over that of the slacks *S*. The problem can be solved by using a two-phase linear programming-based procedure. In the first phase, φ_p is maximised without including the slacks S_r^+ (r=1,…,s) in the model. In the second phase, the optimal value of φ_p, shown by φ_p^*, is replaced in the constraint of model (3) so that the problem is solved with a new objective function which maximises $\sum_{r=1}^{s}S_r^+$ (more details about this two-phase procedure can be found, e.g., in Thanassoulis 2001).

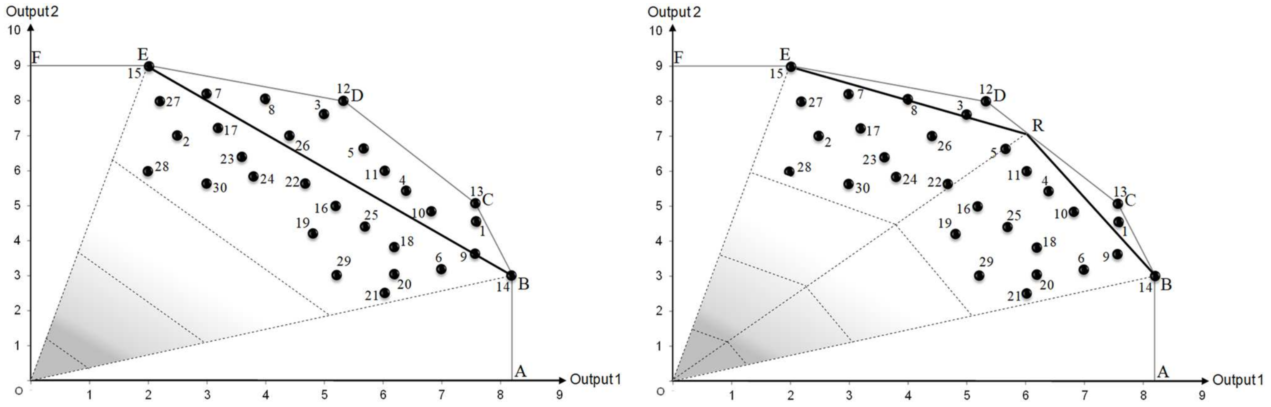
The regression line (still under the ordinary least squares principle) can now be estimated through the units with their new output levels $\tilde{Y}_j=(\tilde{y}_{1j},\tilde{y}_{2j},\ldots,\tilde{y}_{sj})$. In our graphical example, the resulting incentive map is shown by line KL in Fig. 2(b). As can be seen, using projected units in determining the incentive map means that KL is not influenced by any historical inefficiency in the way GH in Fig. 2(a) is. Therefore, it is immune to the problem that inefficiencies are incorporated in determining the incentive maps. In certain cases, however, the method may still require a unit to achieve better performance than that determined as feasible by (2). Taking again DMU$_6$ in Fig. 2(b) as an example, this approach requires the unobserved output levels at 6^* to entitle this DMU to a standard unit of costs. To overcome this problem, Thanassoulis (1996) proposed the following model:

$$\min\left\{\sum_{j=1}^{n}\sum_{r=1}^{s}u_r\tilde{y}_{rj}\left|\sum_{r=1}^{s}u_r\tilde{y}_{rj}\geq c_j\quad j=1,\ldots,n,\qquad u_r\geq 0\quad r=1,\ldots,s\right.\right\}. \tag{4}$$

In this model, u_r (r=1,…,s) are variables. The objective function ensures that the determined incentive map minimises the aggregate cost savings of DMUs subject to the constraint that no DMU is required to achieve a better performance than that determined as feasible by (2). The result of this model, applied to the data of our example, is the line BE depicted in Fig. 3(a). This incentive map is neither influenced by historical inefficiencies nor does it lead to any infeasible plan of cost efficiency savings. Hence, it can be used as benchmark to derive incentive plans for the operating units under central management with minimum degree of decentralisation for DMUs to choose their own output unit costs. Imposing this level of decentralisation, the "incentive map-based cost efficiency" (in the following, it may also be abbreviated as "efficiency") of a DMU$_p$ will be computed by $\theta_p = (u_1^* y_{1p} + u_2^* y_{2p}) / c_p$, in which (u_1^*, u_2^*) are the optimal values in model (4). Accordingly, the units located below line BE (e.g. DMU$_2$) will not be fully compensated and thus encouraged to make cost savings. By contrast, the units on the line (e.g. DMU$_9$) are fully compensated, while units above BE (e.g. DMU$_3$) are overcompensated by being given higher budgets than their projected costs.

There is, however, still a potential major problem with using a single incentive map derived via (4). As can be seen in Fig. 3(a), too many DMUs would be overcompensated resulting in insufficient pressure for the units to improve efficiency. This leads to the need for a hybrid system which is less demanding with respect to cost savings than a fully decentralised one yet more efficient than a fully centralised one of the type BE in Fig. 3(a). The idea of such a hybrid system which allows for different degrees of decentralisation is explained next.

Figure 3. Representation of the proposed incentive maps under two different degrees of decentralisation



(a)

(b)

Let us assume that central management wishes to set two incentive maps, perhaps clustering units by their mix of outputs. For example, in funding police forces, where the mix of type of crime may differ between rural and urban areas, there may be different incentive maps for urban vs. rural police forces. The emerging picture is depicted in Fig. 3(b) as an extension of Fig. 3(a). In Fig. 3(a), when a single incentive map is used, BE is identified as that map, which envelops the efficient projections of DMUs from below in the sense that no part of the DEA efficient frontier is below BE while BE is as close to the efficient frontier as possible. This means that the incentives corresponding to BE require the highest output levels possible in order to entitle a DMU to a standard unit of costs, but subject to requiring no DMU to secure in excess of DEA-efficient output levels for a standard unit of costs. The same principle can now be used to identify the incentive map in each one of the two desired clusters of units. Under the assumption that the central management wishes the incentive map of each cluster to be as close to the efficient frontier as possible (i.e. to have as efficient unit output costs as possible), it is desirable to cluster DMUs with a similar mix of output levels provided that the resulting target is efficient according to (2). Clustering by output mix means DMUs would “agree” on the relative magnitudes of unit output costs, which would maximise the level of compensation they would be entitled to for given output levels.

Thus, for example in Fig 3(b), it is possible to cluster the DMUs to the left of OR whose maximum efficient output levels lie on RE in one cluster and the rest of the DMUs in the second cluster. This defines RE as the incentive map for the first and BR as the incentive map for the second cluster. As both maps lie above BE, the resulting incentives encourage higher cost savings for a given set of output levels than the single cluster with the incentive map BE does. By contrast, these incentives induce lower cost savings than in the decentralised case with three incentives maps BC, CD and DE. It should be noted that the incentive maps need not share the same – real or projected – efficient unit such as R in Fig. 3(b). The number of units under each incentive map and the anchor points on the DEA frontier for each incentive map will be data dependent. If the central management would wish to impose a requirement of a minimum number of units by cluster, this could be incorporated into the model but it will have implications for the level of efficiency savings demanded for some and the incentive levels pertaining to other individual units. Such a model modification may be a fruitful topic for further research but has not been pursued here.

It is worth to contrast the effect of using incentive map-based cost efficiencies and “super-efficiencies” computed respectively by our method and those based on the Andersen and Petersen (1993) approach, developed first by Bogetoft (1997) in incentive regulation (see also Agrell et al. 2005; Bogetoft 2013). In our approach, based on incentive maps, some DEA non-boundary units such as 3 and 1 in Fig. 3(b) will be overcompensated. Whereas all DEA efficient DMUs will at least be fully compensated, some of them will not be overcompensated, such as E and B in Fig. 3(b). Thus, our approach represents a difference in philosophy from the approach originally proposed by Bogetoft (1997). Rather than restricting overcompensation to only originally DEA efficient units, the proposed incentivisation mechanism also offers that opportunity to other units which are close to but not on the DEA efficient boundary. This can lead to revealing efficient practices that may push the efficient frontier beyond the point which originally would have been determined by referring to DEA efficient DMUs only. This comes at the cost from the central management perspective of not being as demanding of efficiency savings as the approach by Bogetoft (1997) may have demanded, based on the criterion of super-efficiency. The benefit, however, is that it raises the number of potential DMUs that can be incentivised to outperform their already strong performance and therefore raises the chances of efficient practices being revealed.

The transition from super-efficiency to incentive maps as the basis for compensation creates winners and losers. Units such as E and B (Fig 3(b)) sitting on an incentive map may have qualified for compensation under super-efficiency but they will not do so now, though they will not be required to make any efficiency savings either. In contrast, units such as C and D which attain output levels in excess of what their incentive map requires will get an incentive benefit proportional to their attainment over and above that required by their incentive map. This may be a higher or lower compensation than might have been the case under super-efficiency. Incentive maps as a criterion is by design based on a cluster of units while super-efficiency is a unit-specific feature. A super-efficient unit can be simply using a rather unusual mix of outputs and so be a self-assessor super-efficient with few, if any, other units having a similar output mix. Units in the extreme ends of an incentive map such as B and E run this risk. On the contrary, units such as C and D with an output mix similar to that of many other units are less likely to be self-assessors. Ultimately, therefore, whether incentive compensation is based on super-efficiency or on incentive maps as proposed here it will lead to winners and losers. Self-assessors will tend to be winners under super-efficiency and losers under an incentive maps scenario. The reverse will be generally the case for efficient units with output mix very similar to that of many other units. Our approach could prove to the disadvantage of efficient units with unusual mixes of outputs. Overall, our approach will temper compensation levels as it will be sharing them amongst a larger number of units than simply those that are efficient.

It is possible to amalgamate the approach recommended here with the one based on super-efficiency so that the compensation is equivalent to the higher of the two compensation levels resulting from incentive maps and the super-efficiency criterion, respectively. In this manner, there will be no losers with regard to the level of compensation due to moving to an incentive maps approach. While this will prove to be more expensive from the central management perspective, the benefit will be that all DEA efficient units are brought into the set that may reveal efficient practices.

3.2. Determining Incentive Maps in the Presence of Varying Degrees of Decentralisation

We now formulate our depicted incentive methodology in mathematical terms. Let n be the number of DMUs of a multi-unit organisation. Further, let us assume that the units enjoy freedom in deciding the relative priorities of output levels and that the central management judges the efficiency of the units with reference to their aggregate operating costs relative to the output levels they deliver. Finally, assume that the central management is willing to consider no more than K sets of unit output costs to take the varying output mixes into account which result from the autonomy units enjoy in setting priorities over output levels. The corresponding clusters and the incentive maps can be obtained by solving the following non-linear mixed integer programming problem:

$$\min\left\{\sum_{j=1}^{n} ER_j \;\middle|\; \begin{array}{lll} \delta_{kj}\sum_{r=1}^{s}\hat{u}_{kr}\tilde{y}_{rj} \le ER_j & j=1,\dots,n, & k=1,\dots,K \\ \delta_{kj}\sum_{r=1}^{s}\hat{u}_{kr}\tilde{y}_{rj} \ge \delta_{kj}\cdot c_j & j=1,\dots,n, & k=1,\dots,K \\ \sum_{k=1}^{K}\delta_{kj}=1 & j=1,\dots,n & \\ \delta_{kj}\in\{0,1\} & j=1,\dots,n, & k=1,\dots,K \\ ER_j \ge 0 & j=1,\dots,n & \\ \hat{u}_{kr}\ge 0 & r=1,\dots,s, & k=1,\dots,K \end{array}\right\}. \tag{5}$$

In this model, there exist *K* clusters whose related constraints have been incorporated into the model by means of δ_{kj} as a set of auxiliary binary variables. When $\delta_{kj}=0$, the corresponding constraints in the first two sets in (5) are inoperative. When $\delta_{kj}=1$, the corresponding constraints in the first set in (5) in conjunction with the minimisation objective function, ensure that ER_j takes the aggregate costs the efficient output levels $\tilde{y}_{rj}$ of DMU$_j$, determined via a DEA model such as (3), would justify under the output unit costs $\hat{u}^*_{kr}$ determined through the solution of (5). The model ensures that for each DMU$_j$ only one δ_{kj} is 1, which in effect assigns DMU$_j$ to the k^{th} incentive map. Thus, for each value of *k* we have a cluster of DMUs which share a given incentive map. The constraint corresponding to $\delta_{kj}=1$ in the second set of constraints in (5) also guarantees that the efficient output levels $\tilde{y}_{rj}$ of DMU$_j$ justify at least its observed aggregate costs c_j. The objective function ensures that the incentive maps derived will minimise the aggregate costs across all DMUs. This in turn means the incentive maps will have associated as efficient unit output costs as possible within the second set of constraints in (5).[3]

It is easy to see that the larger the number of permitted clusters *K*, the lower will be the optimal value of the objective function of (5). In turn, the aggregate costs that the *n* DMUs will collectively justify through their efficient output levels will be lower, and so higher levels of savings will be demanded by the central management from the DMUs. Thus, it is in the interests of DMUs to be assessed with as few clusters as possible, provided the unit output costs used require no DMU to attain savings beyond what its efficient output levels justify under fully decentralised DEA. By contrast, under this constraint, the central management will prefer more incentive maps so that there is more flexibility on unit output costs than those compatible with the least efficient of the efficient units. Thus a conflict of interest is created between the peripheral units and the central management. The model

[3] We note that the proposed model is not designed to balance in some way the number of units under each cluster. However, as stated earlier, this could be done if the management requires it.

developed here provides information for resolving this conflict as the impact on aggregate costs across the system can be evaluated under alternative numbers of clusters of DMUs, i.e. incentive maps.

It should be noted that the incentive maps derived via this model tend to cluster DMUs by the mix of their efficient output levels. This is because the mix of the efficient output levels of a unit determines which ratio of output unit costs would justify a higher aggregate costs for those output levels. Note that the efficient output levels resulting from model (3) largely mirror the mix of the observed output levels of the unit since the model projects first radially the observed output levels of a unit before seeking Pareto efficient levels for those outputs. Clustering by mix of efficient (and in effect observed) output levels is illustrated by Fig. 3(b) in which DMUs in the first cluster, sharing the incentive map RE, offer a larger ratio of output 2 to output 1 than the DMUs in the second cluster sharing the incentive map BR.

Let us now assume that the model in (5) assigns DMU$_j$ to cluster *k*. The unit output costs of cluster *k* will lead to the efficiency score in (6) for DMU$_j$ as

$$\theta_j = \frac{\sum_{r=1}^{s} \hat{u}_{kr}^{*} y_{rj}}{c_j} \tag{6}$$

where $\hat{u}_{kr}^{*}$ (*r*=1,…,*s*) are the costs per unit of output *r* in cluster *k* as determined through model (5). We propose that this θ_j be used within the incentivisation formula in (1).

Thus, inefficient units with $\theta_j < 1$ will not be fully compensated to encourage them to make efficiency savings. In contrast, units with $\theta_j = 1$ are fully compensated while those characterized by $\theta_j > 1$ are rewarded by higher budgets for a period of time than their observed costs c_j. This gives such units an incentive to retain their efficiency because the higher the efficiency score a unit has the larger is the budget to be used by its local management.

It should be noted that the “efficiency” figure θ_j computed in (6) will generally be higher than that which would result for the unit had the traditional DEA model (3) been used, yielding the efficiency $1/\varphi_j$. This is because the unit costs used in (6) by specification require lower output levels to justify the observed costs of at least some of the DEA-efficient units. For example, in Fig. 3(b) all units using the unit costs of incentive map ER will have higher efficiency θ_j than $1/\varphi_j$ apart from unit E which will have the same efficiency $\theta_j = 1/\varphi_j = 1$ under DEA and the formula in (6). Thus, the incentive scheme we propose by using θ_j in the formula in (1) already offers a reduction to the savings that central management would require of DMU$_j$ which is inefficient in DEA terms, over and above the discount on demanded savings inherent in the value of $\rho \leq 1$ that central management adopts within the incentivisation formula (1). The central management needs therefore to be mindful of this dual discount process, the one through $\theta_j \geq 1/\varphi_j$ reflecting the degree of decentralisation that central management is willing to entertain and the one through ρ reflecting potential uncertainties over data and giving units time to adjust operating practices to more efficient modes.

In the context of applying model (5), the central management, perhaps in consultation with the autonomous units, can specify the number of clusters of DMUs to be formed. Under the condition that the unit output costs require no DMU to attain savings beyond what its efficient output levels justify under fully decentralised DEA, the model raises the aggregate cost savings to be delivered by the DMUs the larger the number of clusters K. At one end of the spectrum, when K=1, the binary variables in model (5) are redundant and it reduces to the model in (4). As noted earlier, the efficiencies from this model when used within the incentive formula in (1) by a fully centralised system deliver the lowest level of savings and the maximum level of rewards for the units as a whole. At the other end of the spectrum, when K=n (i.e. each DMU can be a cluster), we have a fully decentralised system and the efficiencies to be used in (1) are those resulting from the classical DEA model in Charnes et al. (1978). These efficiencies deliver the maximum level of savings that could be expected from the system of DMUs and nil rewards. For 1≤K≤n, we have the system being proposed in this paper where the efficiencies in (6), used within the incentive formula (1), will deliver savings and rewards between those resulting from K=1 and K=n. The central management can decide on the number of clusters K appropriate for the level of diversity of operating environments covered by the system of units. This in turn will determine the level of expected savings and granted rewards, respectively.

One potentially important issue for units in our approach is how demanding of efficiency savings are the incentive maps relative to each other. This in turn can be measured by the distance of each incentive map from the original DEA frontier. Model (5) is set up to minimise the difference between the aggregate savings that would be demanded under basic DEA versus the proposed framework where incentive maps are benchmark by cluster. In graphical terms, this is minimising a measure of the distance in aggregate between the DEA frontier and the incentive maps. The shorter the distance the more demanding the incentive map. The location that model (5) will yield for each incentive map is data dependent. If units are of similar scale size, relatively evenly spread across output mixes and of similar relative efficiency, the model would yield incentive maps covering approximately similar numbers of units and of similar distance from the original DEA frontier. The more these uniformities break down in terms of DEA efficiency, scale size and spread across output mixes, the more dissimilar will also be the incentive maps both in terms of the number of units they each cover and, more importantly, their distance from the original DEA frontier. Of course, it is possible to impose restrictions on the distance of each incentive map from the original DEA efficient frontier: This issue could be addressed in further research after perhaps developing a concept of “fairness” or other criteria for setting compensation and efficiency saving levels relative to original DEA efficiencies.

The model in (5) is non-linear with binary variables for which optimal solutions cannot be computed efficiently even for problem instances of small sizes. We now show how this model can be transformed to a mixed integer linear programming (MILP) problem in which one of the hard constraints with binary variables also becomes linear. With the notation as in model (5), the first two sets of constraints can be re-written as follows:

$$\sum_{r=1}^{s}\delta_{kj}\hat{u}_{kr}\tilde{y}_{rj} \le ER_j; \qquad \sum_{r=1}^{s}\delta_{kj}\hat{u}_{kr}\tilde{y}_{rj} \ge \delta_{kj}\cdot c_j, \quad j=1,...,n, \quad k=1,...,K. \tag{7}$$

Considering the right hand side of the second constraint – which has the binary variable δ_{kj} – and the definition of the variables $\hat{u}_{kr}$ in both constrains above, substitute $\delta_{kj}\hat{u}_{kr}$ with a single variable so that $\delta_{kj}\hat{u}_{kr} = u_{kr}$. Hence, these constraints can be transformed to the equivalent constraints

$$\sum_{r=1}^{s}u_{kr}\tilde{y}_{rj} \le ER_j; \qquad \sum_{r=1}^{s}u_{kr}\tilde{y}_{rj} \ge \delta_{kj}\cdot c_j, \quad j=1,...,n, \quad k=1,...,K \tag{8}$$

where u_{kr} are also positive variables. The result is an equivalent MILP problem as follows:

$$\min\left\{\sum_{j=1}^{n}ER_j \;\middle|\; \begin{array}{ll} \sum_{r=1}^{s}u_{kr}\tilde{y}_{rj} \le ER_j & j=1,...,n, \quad k=1,...,K \\ \sum_{r=1}^{s}u_{kr}\tilde{y}_{rj} \ge \delta_{kj}\cdot c_j & j=1,...,n, \quad k=1,...,K \\ \sum_{k=1}^{K}\delta_{kj}=1 & j=1,...,n \\ \delta_{kj}\in\{0,1\} & j=1,...,n, \quad k=1,...,K \\ ER_j \ge 0 & j=1,...,n \\ u_{kr}\ge 0 & r=1,...,s, \quad k=1,...,K \end{array}\right\}. \tag{9}$$

Exact algorithms in commercial software packages such as AIMMS and GAMS can solve the above MILP problem. For example, AIMMS with CPLEX solver applies the branch-and-cut algorithm together with several strategies to speed up the search process in solving MILP problems to optimality. A series of numerical experiments with AIMMS demonstrated that the computational time is very fast even for problems of medium and moderately large sizes, e.g., with 100 DMUs, five input-output variables and three clusters. However, optimal solutions for problems of this type cannot be computed efficiently for large-scale problem instances with probably more demanding combinations of problem parameters, e.g. the one of the German savings banks with 400 DMUs, one input and four outputs where the number of clusters is set to be higher than three. Therefore, heuristic approaches which are able to find near-optimal solutions have to be taken into account.

A comprehensive consideration of heuristic approaches to solve MILP problems is not pursued within this paper but can be found in, e.g., Wolsey (1998). Suffice it to say that a conventional solver – e.g. CPLEX in AIMMS – finds one or more suboptimal solutions in a reasonable computing time (See IBM Knowledge center about CPLEX, IBM 2017). Our own experiments demonstrated that one is able to solve efficiently problems of large sizes – like the case of German savings banks to be discussed in Section 4 – within a pre-specified tolerance level for distance from optimality. In this paper, we apply a heuristic for a *relative optimality tolerance* (*ROT*) to speed up the search process. *ROT* controls the quality of the solution found compared to the solution of the respective relaxed linear programming problem, i.e. $\left|BIS - ORP\right| / BIS < ROT$, where *BIS* is the current best integer solution while *ORP* is the optimal objective value of the relaxed linear programming problem. The branch-and-cut algorithm to solve the MILP problem stops as soon as this relation is satisfied.

3.3. A Dynamic Incentivisation Approach

The external environment of operating units such as government rules and regulations and the economic conditions are likely to change over time, as can the internal environment which can be affected by the goals, internal policies and demands of the organisation’s stakeholders (Afsharian and Ahn 2014). For example, an observed and sustained improvement in the environment of the operating units and in their own productivity may lead to the central management’s expectation that the units should continue to improve their productivity going forward. However, the incentive formula in (1) does not provide for such a dynamic change in productivity. We therefore propose the following modification of the formula in (1) in order to include an expectation of productivity improvement over time:

$$c_p^{q*} = c_p\left(1 - \Omega_p - \psi_p\right)^q \qquad where \qquad \Omega_p = \left[1 - \rho_p(\theta_p - 1)\right]^{1/Q} - 1. \tag{10}$$

Within this dynamic incentive scheme, it is assumed that the central management wishes to provide incentives for DMU$_p$ for each time unit *q*, for a pre-determined regulatory period with a length of *Q* time periods (e.g. years), *q*=1,…,*Q*. Similar to (1), c_p are the observed costs of the operations of DMU$_p$. θ_p is the incentive map-based efficiency of DMU$_p$ during the latest reference time period, which can be computed by our proposed approach in Section 3.2 under a pre-specified number of incentive maps *K*. Hence, with the same definition of the parameter ρ_p as in (1), Ω_p defined in (10) represents the expected efficiency improvement “per year”. The expression for $c_p^{q^*}$ as defined in (10), apart from the annual efficiency savings Ω_p, also includes the parameter ψ_p, which reflects productivity gains that can be expected across all units, including the efficient units from one time period (e.g. year) to the next. ψ_p is traditionally referred to as “X-factor” or “boundary shift” and is deployed extensively in regulation (e.g. see, Bernstein and Sappington 1999; Bogetoft 2013) to reflect sector or industry productivity gains over and above the DMU-specific gains estimated and reflected in Ω_p. Thus, assuming a DMU specific efficiency improvement Ω_p% per year and an annual productivity gain of $100\psi_p$%, $\left(1-\Omega_p-\psi_p\right)^q$ would be the overall fraction of c_p compensated in time period q.

In order to operationalise the incentivisation system based on formula (10), the central management must determine ψ_p. This can be done by analysing historical data over a pre-determined time span consisting of T time periods, e.g., t=1,…,T. We propose a framework next by which one can determine the X-factor under different degrees of decentralisation. Our approach is based on the so-called contemporaneous Malmquist productivity index of Färe et al. (1992). They use DEA to compute the Malmquist index of productivity change and decompose it into technical change (boundary shift) and efficiency change (catch-up) components (for a review of the Malmquist index, see Afsharian and Ahn 2015).

Let n be the number of DMUs, K the clusters (degree of decentralisation), c_j^t the level of costs and $Y_j^t=(y_{1j}^t,y_{2j}^t,...,y_{sj}^t)\in\Re_+^s$ the level of outputs of DMU$_j$ (j=1,…,n) in time period t (t=1,…,T). Therefore, there exist T production possibility sets of feasible input-output combinations as

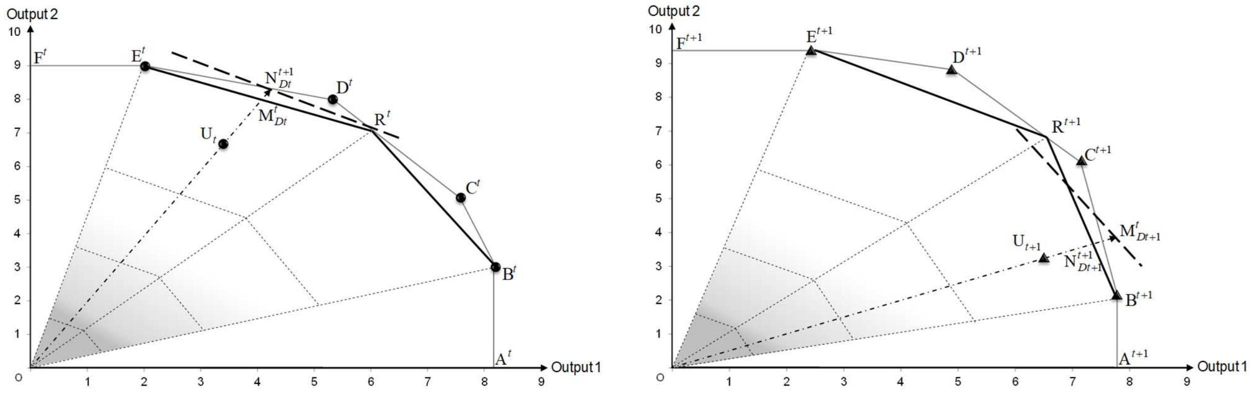
$$PPS^t=\left\{(c^t,Y^t)\in\Re_+\times\Re_+^s \;\middle|\; c^t\geq\sum_{j=1}^{n}\lambda_j^t c_j^t,\quad Y^t\leq\sum_{j=1}^{n}\lambda_j^t y_{rj}^t,\quad \lambda_j^t\geq 0,\quad j=1,...,n\right\} \tag{11}$$

where the respective set of incentive maps in each period t determined by (11) will be denoted here by $IM^t(im_1^t,...,im_K^t)$.

Assuming a constant returns to scale technology, we propose a measure for the efficiency change (EC) and for the technical change (TC) and then combine them to derive a measure of productivity change. It is recalled that the incentive formula in (10) uses incentive maps that are enveloped by the DEA frontier and so the components of the Malmquist index, which use the DEA frontier as computed by Färe et al. (1992), cannot be applied directly within the formula in (10).

Let us first illustrate graphically the way that the efficiency change component can be measured for an individual DMU$_p$ over two time periods t and t+1. In Fig. 4, the production possibility sets for both periods are bounded by ABCDEF whose corner points are denoted with a respective superscript t and t+1. In this example, we assume that the efficiency is measured with a degree of decentralisation of K=2. The two incentive maps in each period $IM^t(im_1^t,im_2^t)$ and $IM^{t+1}(im_1^{t+1},im_2^{t+1})$ are indicated by RE and BR, with the respective superscripts t and t+1. The dotted lines indicate a portion of the boundary in period t+1 (panel a) and t (panel b) where it is relevant for measuring the productivity change for DMU$_p$ under evaluation, represented by U$_t$ and U$_{t+1}$, in periods t and t+1, respectively.

Figure 4. Representation of incentive maps over time



(a) (b)

The efficiency measure of a unit under evaluation in a specific time period is defined with respect to the incentive map to which it is assigned in that period. Therefore, the efficiency of U$_t$ and U$_{t+1}$ will be measured by $\mathrm{OU}_t\ /\ \mathrm{OM}^t_{Dt}$ and $\mathrm{OU}_{t+1}\ /\ \mathrm{ON}^{t+1}_{Dt+1}$, in periods t and t+1, respectively. An efficiency change component indicates whether a unit under evaluation is closer to or further away from a respective incentive map in period t+1 compared to its situation in period t. This can be captured for DMU$_p$ in Fig. 4 by $\left(\mathrm{OU}_{t+1}\ /\ \mathrm{ON}^{t+1}_{Dt+1}\right)/\left(\mathrm{OU}_t\ /\ \mathrm{OM}^t_{Dt}\right)$ or, for the general case, be formulated straightforwardly as

$$EC_p^{t,t+1} = \frac{\theta^{t+1}(c_p^{t+1}, Y_p^{t+1})}{\theta^t(c_p^t, Y_p^t)} \tag{12}$$

where $\theta^t(c_p^t, Y_p^t)$ and $\theta^{t+1}(c_p^{t+1}, Y_p^{t+1})$ represent the efficiency score of DMU$_p$ (p=1,...,n) in period t and t+1, respectively, relative to the incentive map for the unit at the time period concerned.

Consider U$_t$ in Fig. 4(a) now. Its incentive map in period t, and the incentive map for its cluster in period t+1 (dotted line) are shown. Similarly, the incentive map for U$_{t+1}$ for period t+1 and the incentive map for period t (dotted line) are shown in Fig. 4(b). In Fig. 4(a), the projection of U$_t$ on its incentive map in period t, indicated by M^t_{Dt}, moves to N^{t+1}_{Dt} on the incentive map in period t+1. Thus, the technical change for U$_t$ can be measured by $\mathrm{ON}^{t+1}_{Dt}\ /\ \mathrm{OM}^t_{Dt}$. This is equivalent to $(\mathrm{OU}_t\ /\ \mathrm{OM}^t_{Dt})\ /\ (\mathrm{OU}_t\ /\ \mathrm{ON}^{t+1}_{Dt})$, which represents the ratio of the efficiency scores of DMU$_p$ in period t according to the incentive map RE in period t and period t+1, respectively. Similarly, the technical change for U$_{t+1}$ in Fig. 4(b) can be measured by $\mathrm{ON}^{t+1}_{Dt+1}\ /\ \mathrm{OM}^t_{Dt+1}$, which represents the ratio of the efficiency scores of DMU$_p$ in period t+1 according to the incentive map BR in period t and period t+1, respectively, i.e. $(\mathrm{OU}_{t+1} / \mathrm{OM}^{t}_{Dt+1}) / (\mathrm{OU}_{t+1} / \mathrm{ON}^{t+1}_{Dt+1})$. With respect to these captured changes, we can define the individual technical change component of DMU$_p$ over two time periods t and t+1 as the geometric mean of the above measures, i.e. as $TC_p^{t,t+1} = \sqrt{(\mathrm{OU}_t / \mathrm{OM}^{t}_{Dt}) / (\mathrm{OU}_t / \mathrm{ON}^{t+1}_{Dt}) \times (\mathrm{OU}_{t+1} / \mathrm{OM}^{t}_{Dt+1}) / (\mathrm{OU}_{t+1} / \mathrm{ON}^{t+1}_{Dt+1})}$. This can be formulated for the general case as

$$TC_p^{t,t+1} = \left[\frac{\theta^t(c_p^t, Y_p^t)}{\theta^{t+1}(c_p^t, Y_p^t)} \times \frac{\theta^t(c_p^{t+1}, Y_p^{t+1})}{\theta^{t+1}(c_p^{t+1}, Y_p^{t+1})} \right]^{\frac{1}{2}} \tag{13}$$

where $\theta^z(c_p^w, Y_p^w)$, $z, w = t, t+1$ represent the efficiency score of DMU$_p$ (p=1,...,n) observed in period w (w=t, t+1) against a respective incentive map $im_{\bar{k}}^z$ ($\bar{k} \in \{1,...,K\}$) in period z (z=t, t+1).

Having defined EC and TC, the new framework of the Malmquist index (MI_{im}) combines these components to measure productivity change over the two time periods t and t+1 as $MI_{im}(c_p^t, Y_p^t; c_p^{t+1}, Y_p^{t+1}) = EC_p^{t,t+1} \cdot TC_p^{t,t+1}$. After substitutions and algebraic manipulations, the following expression for the Malmquist index is derived:

$$MI_{im}(c_p^t, Y_p^t; c_p^{t+1}, Y_p^{t+1}) = EC_p^{t,t+1} \cdot TC_p^{t,t+1} = \left[\frac{\theta^t(c_p^{t+1}, Y_p^{t+1})}{\theta^t(c_p^t, Y_p^t)} \times \frac{\theta^{t+1}(c_p^{t+1}, Y_p^{t+1})}{\theta^{t+1}(c_p^t, Y_p^t)} \right]^{\frac{1}{2}}. \tag{14}$$

A value of this Malmquist index or any of its components less than one denotes regress, while a value greater than one implies progress (e.g., if MI=1.05 it means the same costs deliver 5% more output in period t+1 compared to period t and the reverse if say MI=0.95). In addition, a value of one indicates unchanged productivity between periods t and t+1 by the DMU concerned. The expression in (14) is essentially the traditional Malmquist index as defined by Färe et al. (1992) and reproduced by Thanassoulis (2001), among others. The key difference is that the traditional Malmquist index computes the efficiencies with reference to the boundary of the PPS, whereas we use the incentive maps for reference. Further, the boundary shift captured is that of incentive maps rather than of the PPS frontier (i.e. the traditional Malmquist index is a special case of our approach with K=n, where the incentive maps will be the frontier of the PPS). Hence, with an appropriate degree of decentralisation K imposed by central management, our approach can not only overcome the problem of excessive rents but also reduce the risk of failure inside the system due to an overestimated productivity gain.

Using the expressions in (12), (13) and (14), the proposed Malmquist index as well as its components for DMU$_p$ (p = 1,…,n) over time periods t and t+1 can be computed by means of four measures of efficiency as $\theta^z(c_p^w, Y_p^w)$, z, w=t, t+1. After having solved model (9) in each period of time and having found the incentive maps

$IM^t(im_1^t,...,im_K^t)$ and $IM^{t+1}(im_1^{t+1},...,im_K^{t+1})$, we can compute $\theta^t(c_p^t,Y_p^t)$ and $\theta^{t+1}(c_p^{t+1},Y_p^{t+1})$ straightforwardly by means of formula (6). The computation of $\theta^t(c_p^{t+1},Y_p^{t+1})$ and $\theta^{t+1}(c_p^t,Y_p^t)$ as cross-period efficiency measures can also be done by considering a respective cross-period incentive map for DMU$_p$. It is logical to use, e.g., for the input-output levels of DMU$_p$ observed in period t+1 the period t cluster that would offer the highest efficiency rating for DMU$_p$ with reference to its t+1 input-output levels. Thus, $\theta^t(c_p^{t+1},Y_p^{t+1})$ can be computed by using the following expression:

$$\theta^t(c_p^{t+1},Y_p^{t+1}) = \max\left(\frac{\sum_{r=1}^{s} u_{1r}^{t*} y_{rp}^{t+1}}{c_p^{t+1}}, \frac{\sum_{r=1}^{s} u_{2r}^{t*} y_{rp}^{t+1}}{c_p^{t+1}}, ..., \frac{\sum_{r=1}^{s} u_{Kr}^{t*} y_{rp}^{t+1}}{c_p^{t+1}}\right). \tag{15}$$

$\theta^{t+1}(c_p^t,Y_p^t)$ can be computed in a similar manner.

As outlined earlier, in the context of the incentivisation regime, central management would normally use the formula in (10) to set budgets for the semi-autonomous units over a regulatory period with the length of Q. Thus, while c_p within (10) would often refer to observed operating expenditures during the latest complete time period, ψ_p will likely be based on some average (e.g. the geometric mean) of the chain $TC_p^{t,t+1}$ values observed in the preceding T>1 time periods. The value of ψ_p used in (10) would then be intended to reflect the expectation of mean annual productivity gain going forward.

4. An Empirical Illustration Using Data from German Savings Banks

We refer to the German savings banks whose structure has already been outlined in order to illustrate the incentivisation approach developed in this paper. We begin with the identification of the input and output factors used for measuring the efficiency of the banks, adopting the production perspective (see, e.g., Berger and Humphrey 1997) of banking. This places emphasis on how banks as providers of products and services create their outputs by using a minimum level of resources (an extensive literature review can be found, e.g., in Ahn and Le 2014). It should be emphasized that the empirical application in this section is only illustrative of our approach and does not necessarily reflect the official policy of the group of German savings banks.

We have specified a single input, namely *total operating expenses*. The corresponding outputs are *total deposits*, *net loans*, *other earning assets* and *total non-interest operating income*. The outputs cover most of the products and services offered by savings banks. Considering *deposits* as output is consistent with the production perspective. It is a proxy for customer services offered by a bank. *Net loans* comprise total loans less reserves for non-performing loans. *Other earning assets* also include total securities, loans and advances to banks. *Non-interest income* is considered as an output which serves as proxy for fee-based products and services provided by the banks. The data have been extracted from the Bankscope database and relate to the 416 savings banks overseen by the central authority DSGV as described in the introduction. The data cover the time period 2010-2014. A total of 16 banks were excluded from the analysis because of unreliable information. In order to make monetary values in different years comparable, the data have been adjusted for inflation. More specifically, by means of the retail price index, the values have been adjusted to 2014 prices.

The proposed models have been formulated under a CRS assumption. In a context where there is asymmetry of information between the central management and the operating units, notably in the regulation of utilities in Europe, the regulator (central management) normally adopts a CRS assumption in assessing the scope for efficiency savings irrespective of the actual returns to scale prevailing locally at individual operating units. This is in order to incentivise them to move to a most productive scale size. The counter argument is of course that there may be costs and other impediments to a unit exploiting economies of scale. Nevertheless, in the spirit of incentive regulation that forms the background to this paper, we assume for our illustration that the central management would adopt a CRS assumption in assessing the scope for savings at the banks which are used here as example.

In order to run the incentive formula proposed in this study, we need to specify the number of clusters *K*, reflecting the level of decentralisation that DSGV (the central authority) would be willing to grant to the savings banks. Since the primary goal of this empirical application is to illustrate the proposed incentivisation framework, we consider a number of alternative values of *K* to illustrate the effects as *K* increases. We begin from the level of decentralisation with a minimum of one incentive map (i.e. *K*=1) and increase the incentive maps up to a threshold at which the level of rewards offered to operating units in the system, denoted by R_k^+, becomes on average less than 1.5% of the aggregate observed expenditure by the corresponding banks. In addition, the results concerning the maximum number of incentive maps (i.e. potentially *K*=400, the classical DEA model) are also reported.

The required instances of the mathematical programming model in (9) were encoded in AIMMS, version 4.14. Within AIMMS, we also applied a heuristic for a *ROT* to speed up the search process (see Section 3.2). *ROT* was set at 0.01. At this tolerance level, the final integer solution deviates from the optimal value by a maximum of 1%.

With respect to the theoretical arguments put forward in Section 3, our proposed model in (9) is expected to raise the aggregate cost savings to be delivered by the DMUs, the larger the number of incentive maps or clusters *K*. For the extreme case of *K*=1, the binary variables in that model are redundant and the model reduces to the one in (4). In this scenario of a central management with minimum degree of decentralisation, the units are grouped into a single cluster so that incentives are determined in a uniform way. As the number of clusters increases, the model in (5) should reduce further the aggregate costs justified by the outputs delivered by the system of banks as a whole. At the other extreme with *K*=400, which represents complete autonomy for each bank and in a sense maximum degree of decentralisation, each DMU is permitted to be its own cluster. Thus, model (5) is expected to generate the same aggregate cost savings as the decentralised DEA model of Charnes et al. (1978) does. The results of our analysis are summarised in Table 1.

Table 1. Results for different degrees of decentralisation

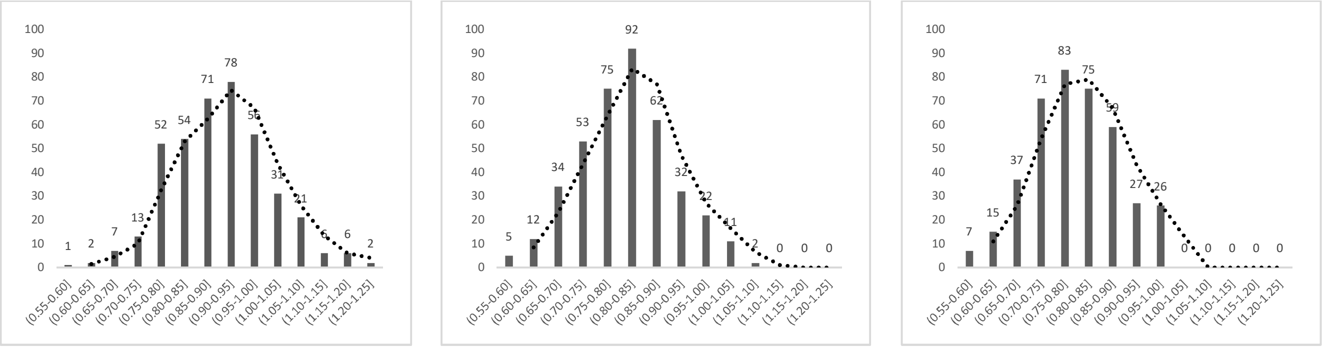
	Degree of decentralisation					
	Minimum	Hybrid				Maximum
	K=1	*K*=2	*K*=3	*K*=4	*K*=5	*K*=400
			Number of banks			
Eff > 1	66 (16.50%)	26 (6.50%)	13 (3.25%)	10 (2.50%)	7 (1.75%)	0 (0.00%)
Cluster 1	400	244	159	107	119	
Cluster 2	–	156	137	106	86	
Cluster 3	–	–	104	94	77	
Cluster 4	–	–	–	93	76	
Cluster 5	–	–	–	–	42	
			Efficiency score			
1st quartile	0.827	0.762	0.746	0.740	0.739	0.733
Median	0.900	0.834	0.811	0.806	0.800	0.793
3rd quartile	0.965	0.903	0.877	0.868	0.867	0.858
			Compensation			
$Av.C^{-}$ (Eff < 1)	0.133	0.177	0.198	0.202	0.203	0.211
$Av.R^{+}$ (Eff > 1)	0.068	0.049	0.031	0.018	0.014	–

As can be seen in Table 1, the level of cost savings on average demanded of inefficient units (those operating units with Eff < 1), represented by $Av.C^{-}$, increases as the number of clusters *K* rises. Thus, under standard DEA (*K*=400), on average, 21.1% of costs of the inefficient banks should be saved while under *K*=1 only 13.3% need to be saved. The opposite trend can be seen for $Av.R^{+}$, which captures the level of rewards on average offered to super-efficient units (those operating units with Eff > 1). Thus, for the case of *K*=1, $Av.R^{+}$ shows rewards of 6.8% given on average to super-efficient units, but this drops to 1.4% (with *K*=5) and to zero for *K*=400 (standard DEA).

Table 1 also shows quartiles of the efficiency scores. Under a single incentive map, 66 banks have an efficiency score greater than one. These banks would be incentivised by higher budgets than their projected costs. More than 75% of the banks (i.e. 300 banks) have an efficiency score greater than 96.5%. These "efficiency" levels testify to the fact that using a single cluster would not lead to budgets that are challenging enough for inducing efficiency in the majority of banks if we maintain that no bank can be asked to attain higher outputs than are DEA efficient for its input costs. By contrast, under the other extreme with the maximum degree of decentralisation, (400 potential clusters) 50% of the banks (i.e. 200 banks) receive an efficiency score less than 79.3%. Moreover, the banks only get non-positive incentives, i.e. the number of banks with an efficiency score greater than one is zero. Therefore, under this maximum number of incentive maps, no opportunity is provided to overcompensating efficient banks in order to improve the overall performance of the system.

The particular shortcoming of both extreme cases can be mitigated by choosing one of the other levels of decentralisation. For the case of, e.g., *K*=3, the efficiency score of more than 50% of the banks is lower than 81.1%, while only 3.25% of the banks (i.e. 13 banks) are awarded with an efficiency score greater than one. Thus, DSGV would demand expenditure savings of 20% or more from 50% of the banks and reward no more than 3.25% of the banks with excess budgets on average of 3.1% (note that these savings and rewards are resulting before ρ is applied by the central management within the formula in (1) to determine the incentivisation power). Fig. 6 compares this case of *K*=3 with the two extremes in greater detail, illustrating the respective distributions of the banks over their efficiency scores.

Figure 6. Distributions of the banks over their efficiency scores



Minimum degree of decentralisation (*K*=1) Decentralisation degree (*K*=3) Maximum degree of decentralisation (*K*=400)

Under minimum degree of decentralisation, the three greatest frequencies occur between the efficiency scores of 0.85 and 1.00, covering 205 banks. Under maximum degree of decentralisation, the three greatest frequencies are located between efficiency scores of 0.70 and 0.85, in regards to 229 banks. As a compromise, *K*=3 places the three greatest frequencies between the scores of 0.75 and 0.90, comprising also 229 banks.

Comparing the frequencies in Fig. 6 and taking into account the level of cost savings and rewards given under different degrees of decentralisation, let us assume that *K*=3 is chosen by DSGV. While we cannot present the results of all individual banks here, Table 2 lists the results of 16 selected banks whose efficiency scores resulting from our approach are greater than or equal to 1. For the sake of comparison, the table also lists the standard DEA efficiency scores as well as the corresponding DEA super-efficiency scores computed by the Andersen and Petersen (1993) approach, adapted by Bogetoft (1997) in the context of incentive regulation.

Table 2. Results of individual efficiencies

Bank	Efficiency under the degree of *K*=3	Standard DEA efficiency	DEA super-effciency
B_{09}	1.040	1.000	1.052
B_{41}	1.005	0.984	0.984
B_{46}	1.041	1.000	1.012
B_{69}	1.000	1.000	1.000
B_{75}	1.041	1.000	1.014
B_{82}	1.046	1.000	1.077
B_{91}	1.020	1.000	1.009
B_{110}	1.000	1.000	1.070
B_{117}	1.000	1.000	1.018
B_{126}	1.058	1.000	1.083
B_{141}	1.018	1.000	1.004
B_{144}	1.043	1.000	1.005
B_{160}	1.002	0.984	0.984
B_{235}	1.055	1.000	1.026
B_{339}	1.002	0.964	0.964
B_{356}	1.032	1.000	1.065
		Compensation	
All 400 banks	81.35%	79.83%	79.95%

The examples in Table 2 show the effects of choosing between the respective approaches. For instance, the efficiency score of bank B_{46} would have been 1.2% above one if the DEA super-efficiency had been chosen. Under the maximum degree of decentralisation (standard DEA), this bank would only have been recognised as fully efficient, losing 1.2% of its potential level of reward. However, this bank has been shown to qualify for overcompensation of 4.1% within our approach (i.e. the degree of decentralisation with *K*=3). As discussed in Section 3.1, this bank is a winner in this transition from DEA super-efficiency to incentive maps-based efficiency as a criterion for compensation. The same phenomenon can be seen for the banks B_{75}, B_{91}, B_{141}, B_{144} and B_{235} whose (super-)efficiency scores would have decreased if either the standard DEA (i.e. the maximum degree of decentralisation) or the DEA super-efficiency had been applied.

In contrast, there are those banks (i.e. B_{09}, B_{82}, B_{110}, B_{117}, B_{126}, B_{356}) which are losers in the sense that their incentive map-based cost efficiencies are less than their respective DEA super-efficiency scores. This explains how the self-assessors tend to be winners under DEA super-efficiency and losers under our incentive map-based approach. We also note that the banks B_{41}, B_{160} and B_{339} are offered rewards in excess of expenditure under *K*=3, wheras they would have been required to save costs, if their standard DEA (in)efficiency scores had been taken into account. This shows how the proposed incentivisation mechanism offers rewards to banks which are close to but not on the standard DEA efficient boundary. An effect of such an approach is also reflected in the mean compensation level to all 400 banks. Using *K*=3 as the degree of decentralisation, 81.35% of the observed costs are compensated, corresponding to demanded savings of 18.65% of observed costs. In contrast, under DEA super-efficiency (which also offers overcompensation), this amount compensated decreases from 81.35% to 79.95% of observed costs. This comparison demonstrates how the proposed approach may incentivise already good performers (but not necessarily only those banks sitting on the efficient boundary under standard DEA) in order to raise the chances of efficient practices being revealed in future.

It is interesting to compare incentive maps as to whether some are more demanding than others in terms of required savings. This can go back to the issue of "fairness" of the incentives system as might be perceived by operating units. The issue will be dependent on the number of clusters used and on the performance of units by their mix of outputs. We demonstrate in Table 3 the comparative savings required by cluster when incentive maps range from *K*=2 to *K*=4.

Table 3. Incentive Rewards and Savings Demanded

	Degree of decentralisation								
	K=2		*K*=3			*K*=4			
	Clusters								
	C_1	C_2	C_1	C_2	C_3	C_1	C_2	C_3	C_4
Total number of banks	244	156	159	137	104	107	106	94	93
Top 10%	103.82	100.24	98.66	97.89	99.22	99.11	96.34	96.67	96.49
Bottom 10%	68.28	64.90	64.38	64.49	67.97	66.81	65.55	64.68	63.30
All banks	86.19	81.26	80.77	81.72	81.99	80.13	81.59	80.09	81.53

As can be seen in Table 3, the banks which belong to the top 10% on efficiency in cluster C_1 under *K*=2 (i.e. 25 banks) are offered 3.82% reward in addition to their observed expenditure (i.e. an incentive budget of 3.82% above their reported costs).[4] The top 10% of units on efficiency in cluster C_2 under *K*=2 are offered only a minor reward of 0.24% of observed expenditure. In contrast, only 68.28% of the observed costs of the bottom 10% of the banks in cluster C_1 is reimbursed, representing demanded savings of 31.72% of observed expenditure. The corresponding figure for the bottom 10% of units in cluster C_2 under *K*=2 is 35.1%.[5] Thus, for *K*=2, we have obvious disparities between clusters on the savings demanded and rewards offered. As the value of *K* increases, these disparities reduce. For *K*=3, the savings demanded range from 2.11% (*K*=3, C_2) to 0.78% (*K*=3, C_3) for the top 10% of units on efficiency in the clusters. Similarly, the savings demanded of the bottom 10% of units on efficiency in each cluster for *K*=3 range from 35.62% (C_1) to 32.03% (C_3) of observed expenditure. The savings demanded across all units of each cluster for *K*=3 range from 19.23% (*K*=3, C_1) to 18.01% (*K*=3, C_3) of observed expenditure, which is very narrow. The differences by cluster further decrease if we adopt *K*=4 (four clusters).

As noted earlier, these differences in levels of compensation and savings demanded by cluster are data dependent, so no generalisable statement can be made. It is possible to surmise that the differences will be lower the more even the distribution of the number of units by cluster is, as indicated in Table 3. This could in turn be because

[4] These savings and rewards are resulting before ρ is applied by the central management within the incentive formula in (1) to determine the incentivisation power.

[5] We should note that these amounts would normally be attenuated via ρ in the incentive formula in (1) and would also be split over the years of the regulatory period with the proposed dynamic incentive scheme in (10).

each subset of units has a similar variance of underlying efficiency across the units. If there are large disparities by cluster and this is deemed an important issue of fairness, as noted earlier, the model used can be adapted to restrict distances of incentive maps for the efficient frontier.

The results in the above tables, relate to the case where no expectation of productivity change over time is incorporated in the incentive formula. However, DSGV may wish to incorporate in the incentive formula its beliefs about potential improvements in productivity that all the banks should be able to achieve over and above any catch up in bank-specific efficiency. In such a case, DSGV would first determine the level of productivity gain that all banks must achieve, i.e. the so called X-factor. This can be done for example by analysing historical data of efficient boundary shift over a pre-determined regulatory period, e.g. 2010-2014 and/or information about productivity gains in financial services within the country. The respective statistics of the Malmquist index and its components for German savings banks are given in Table 4.

Table 4. Results of the Malmquist index and its components

		EC	*TC*	MI_{im}			*EC*	*TC*	MI_{im}
	Min	0.801	0.947	0.830		Min	0.805	0.954	0.868
Adj-period 1:	Max	1.169	1.182	1.217	Adj-period 3:	Max	1.229	1.136	1.308
2010-2011	Mean	0.987	1.055	1.041	2012-2013	Mean	0.996	1.088	1.084
	St. Dev	0.057	0.031	0.058		St. Dev	0.066	0.026	0.069
	Min	0.740	1.036	0.836		Min	0.714	0.882	0.793
Adj-period 2:	Max	1.221	1.317	1.325	Adj-period 4:	Max	1.287	1.116	1.267
2011-2012	Mean	0.946	1.089	1.031	2013-2014	Mean	1.040	1.036	1.077
	St. Dev	0.067	0.036	0.062		St. Dev	0.077	0.042	0.068

As can be seen in Table 4, the mean value (calculated using a geometric mean) of the Malmquist index (MI_{im}) for each of the four adjacent periods (hereafter adj-periods) is greater than 1, signifying that productivity has steadily increased during the analyzed period. On the contrary, given the decomposition of MI_{im} into efficiency change (*EC*) and technical change (*TC*), the average efficiency of the banks has decreased in the first three adj-periods, i.e. the *EC* component reveals negative changes of -1.3%, -5.4% and -0.4%, respectively. Nevertheless, a closer look at the results of the *TC* component highlights its remarkable positive effect on the growth in productivity, i.e. changes over the three first adj-periods are 5.0%, 8.9% and 8.8%, respectively. Therefore, although the efficiency has decreased in these three adj-periods, a positive change in the technology has led to a productivity growth. In the last adj-period, both *EC* and *TC* show positive changes, explaining straightforwardly the productivity growth of 7.7%.

From the results in Table 4, it can be concluded that the savings banks studied have been generally able to improve their performance over time. This has been mainly due to the positive shift in technology, i.e. the benchmark savings banks produced the same level of outputs at lower costs within the analysed time frame. The geometric mean of productivity change over the four years in Table 4 is just under 6%. DSGV can use this as the X-factor to be applied in formula (10). However, there may also be good reasons to adjust the computed X-factor to incorporate aspects like the costs of achieving improvement in the quality of services or transient factors affecting productivity that may not hold in future.

In order to illustrate how an individual X-factor and an individual cost reduction can be imposed, let us assume that DSGV applies a regulatory period of four years. Consider now an inefficient bank whose efficiency θ_p has been captured by our approach as 80%. Hence, according to formula (10), if ρ is set to 0.6, the individual cost reduction per year shown by Ω_p would be 2.9%, i.e. $0.0287 = [1-0.6(0.8-1)]^{1/4} - 1$. Let us suppose that DSGV sets the banks' collective annual X-factor ψ_p to be 3%. Then the cost reduction per year to be used in formula (10) will be 5.9%. As can be seen here, this amount depends on the parameter ρ, the length of the regulatory period Q, the level of decentralisation K and also the way the X-factor ψ is specified. To sum up, the approach proposed in this section should not be applied in a purely mathematical manner but as an initial guideline to properly set the parameters of the incentive formula for a particular unit.

5. Conclusion

In centrally managed multi-unit organisations, the central management differs substantially in the degree to which it controls the day-to-day business of the operating units and imposes policies on them. Operating units under different degrees of decentralisation are given different levels of autonomy in terms of how, e.g., they can take actions, make decisions and cooperate with others to deliver products and services to customers. The central management's challenge is to incentivise the local operating units by means of an appropriate efficiency measurement system which should be in line with the degree of decentralisation. The central management wants to ensure that inefficient units are required to become more efficient and those with good efficiency are incentivised to improve further still. This has to be done in the presence of information asymmetry including the adverse selection issue (i.e. operating units have better information than central management about local available efficient practices), and the moral hazard problem (i.e. there exist conflicts of interest as to which actions should be taken by operating units locally).

In this paper, we propose a DEA-based incentive mechanism that central management can use to meet the objectives of improving efficiency for a system of operating units on which it has varying degrees of control. In order to measure efficiency an adaptation of the approach introduced by Thanassoulis (1996) is first applied to cluster operating units by their output mix. A set of common weights (e.g. cost levels per unit of outputs) is then determined which makes it possible to incentivise the operating units in a manner which most closely reflects their particular operating priorities and environment. The proposed method provides a tool for the central management to decide about the level of cost savings required from the inefficient units and the rewards given to units with an exceptional efficiency. An extension of the method also allows the central management to determine incentive levels recognising that units may experience productivity changes over time. This is done by means of a new framework for computing Malmquist indices to capture productivity gains over time in a decentralised context. Our approach has been illustrated using data from the group of German savings banks.

Besides the fact that central management has control as to how the parameters of the proposed incentive mechanism are specified, two further aspects concerning the choice of the X-factor should be taken into account: First, some additional factors may also be incorporated such as "improvement in quality of services" (see, e.g., Maziotis et al. 2016) or "major structural changes in the industry" (see, e.g., Bernstein and Sappington 1999) in the estimation of the X-factor. Second, although many applications determine a single X-factor to be applied to all units equally, we suggest within our approach that central management should impose multiple or individual X-factors in the system. The reason is that if the X-factor is imposed uniformly across all units, there might be operating units which may earn excessive budgets and thereby jeopardise support for the implementation of the incentivisation system in the group. On the other hand, this single estimated X-factor might be too demanding for some other units, threatening the financial integrity of these units in the system (for more discussions about this issue, see also Bernstein and Sappington 1999). Put another way, boundary shift may be different for units which differ in mix of inputs they use and or unit input prices they face.

Acknowledgement

Funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – AH 90/5-2.

References

Afsharian, M., Ahn, H. 2014. The Luenberger indicator and directions of measurement: A bottoms-up approach with an empirical illustration to German savings banks. *Int. J. Prod. Res.* 52(20), 6216-6233.

Afsharian, M., Ahn, H. 2015. The overall Malmquist index: A new approach for measuring productivity changes over time. *Ann. Oper. Res.* 226(1), 1-27.

Afsharian, M., Ahn, H. 2016. Multi-period productivity measurement under centralized management with an empirical illustration to German saving banks. *OR Spectrum.* 39(3), 881-911.

Afsharian, M., Ahn, H., Thanassoulis, E. 2017. A DEA-based incentives system for centrally managed multi-unit organisations. *Eur. J. Oper. Res.* 259(2), 587-598.

Agrell, P. J., Bogetoft, P. 2001. Should health regulators use DEA. Coordinacion e Incentivos en Sanidad, Asociasion de Economia de la Salud, Barcelona, 133-154.

Agrell, P. J., Bogetoft, P., Tind, J. 2005. DEA and dynamic yardstick competition in Scandinavian electricity distribution. *J. Prod. Anal.* 23(2), 173-201.

Ahn, H., Le, M. H. 2014. An insight into the specification of the input-output set for DEA-based bank efficiency measurement. *Manag. Rev. Quart.* 64(1), 3-37.

Asmild, M., Paradi, J. C., Pastor, J. T. 2009. Centralized resource allocation BCC models. *Omega.* 37(1), 40-49.

Andersen, P., Petersen, N. C. 1993. A procedure for ranking efficient units in data envelopment analysis. *Management science*. 39(10), 1261-1264.

Bernstein, J. I., Sappington, D. E. 1999. Setting the X factor in price-cap regulation plans. *J. Regul. Econ.* 16(1), 5-26.

Berger, A. N., Humphrey, D. B. 1997. Efficiency of financial institutions: International survey and directions for future research. *Eur. J. Oper. Res.* 98(2), 175-212.

Bogetoft, P. 1994. Incentive efficient production frontiers: An agency perspective on DEA. *Manag. Sci.* 40(8), 959-968.

Bogetoft, P. 1997. DEA-based yardstick competition: The optimality of best practice regulation. *Ann. Oper. Res.* 73, 277-298.

Bogetoft, P. 2013. *Performance Benchmarking: Measuring and Managing Performance.* Springer, New York.

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

Cook, W. D., Zhu, J. 2007. Within-group common weights in DEA: An analysis of power plant efficiency. *Eur. J. Oper. Res.* 178(1), 207-216.

Deutscher Sparkassen- und Giroverband (DSGV): Inside the savings banks finance group. 2017a. Accessed December 20, 2017. http://www.dsgv.de/en/facts/publications.html

Deutscher Sparkassen- und Giroverband (DSGV): Missions and objectives. 2017b. Accessed December 20, 2017. http://www.dsgv.de/en/about-us/index.html

Fang, L. 2013. A generalized DEA model for centralized resource allocation. *Eur. J. Oper. Res.* 228(2), 405-412.

Färe, R., Grosskopf, S., Lindgren, B., Roos, P. 1992. Productivity developments in Swedish pharmacies: A non-parametric Malmquist approach. *J. Prod. Anal.* 3, 85-101.

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

Hawdon, D. 2003. Efficiency, performance and regulation of the international gas industry: A bootstrap DEA approach. *Energy Policy.* 31(11), 1167-1178.

IBM: IBM Knowledge center. 2017. http://www.ibm.com/support/knowledgecenter. Accessed December 13, 2017.

Kao, C., Hung, H. T. 2005. Data Envelopment Analysis with common weights: The compromise solution approach. *J. Oper. Res. Soc.* 56(10), 1196-1203.

Kuosmanen, T., Johnson, A., Saastamoinen, A. (2015). *Stochastic nonparametric approach to efficiency analysis: A unified framework. In Data envelopment analysis (pp. 191-244).* Springer, US.

Lozano, S., Villa, G. 2004. Centralized resource allocation using Data Envelopment Analysis. *J. Prod. Anal.* 22(1-2), 143-161.

Mar-Molinero, C., Prior, D., Segovia, M. M., Portillo, F. 2014. On centralized resource utilization and its reallocation by using DEA. *Ann. Oper. Res.* 221(1), 273-283.

Maziotis, A., Saal, D. S., Thanassoulis, E., Molinos-Senante, M. 2016. Price-cap regulation in the English and Welsh water industry: A proposal for measuring productivity performance. *Util. Policy.* 41, 22-30

Roll, Y., Cook, W. D., Golany, B. 1991. Controlling factor weights in Data Envelopment Analysis. *IIE Trans.* 23(1), 2-9.

Roll, Y., Golany, B. 1993. Alternate methods of treating factor weights in DEA. *Omega.* 21(1), 99-109.

Shleifer, A. 1985. A theory of yardstick competition. *RAND J. Econ.* 319-327.

Simpson, C. V. J. 2013. *The German Sparkassen (savings banks): A commentary and case study*. Civitas, London.

Thanassoulis, E. 1996. A Data Envelopment Analysis approach to clustering operating units for resource allocation purposes. *Omega*. 24(4), 463-476.

Thanassoulis, E. 2000. DEA and its use in the regulation of water companies. *European Journal of Operational Research*. 127(1), 1-13.

Thanassoulis, E. 2001. *Introduction to the Theory and Application of Data Envelopment Analysis*. Springer, New York.

Varmaz, A., Varwig, A., Poddig, T. 2013. Centralized resource planning and yardstick competition. *Omega*. 41(1), 112-118.

Vitols, S. 1995. German banks and the modernization of the small firm sector: Long-term finance in comparative perspective. Discussion Paper FS I 95-309. Wissenschaftszentrum Berlin für Sozialforschung.

Wolsey, L. A. 1998. *Integer Programming*. Wiley-Interscience Publication, New York.