Measuring Potential Sub-unit Efficiency to Counter the Aggregation Bias in Benchmarking

Ahn, Heinz; Bogetoft, Peter; Lopes, Ana

Link to publication in CBS Research Portal

Measuring Potential Sub-unit Efficiency to Counter the Aggregation Bias in Benchmarking

Heinz Ahn, Peter Bogetoft, Ana Lopes

Journal article (Accepted version*)

Uploaded to CBS Research Portal: August 2019

Measuring *Potential Sub-Unit Efficiency* to Counter the Aggregation Bias in Benchmarking

Heinz Ahn[a], Peter Bogetoft[b] and Ana Lopes[c]

[a] Corresponding Author. Technische Universität Braunschweig, Institute of Management Control and Business Accounting, Fallersleber-Tor-Wall 23, D-83100 Braunschweig, Germany. Phone: +49 531 391-3610. Email: hw.ahn@tu-braunschweig.de.

[b] Copenhagen Business School, Department of Economics, Porcelaenshaven 16 A, DK-2000 Frederiksberg, Denmark. Phone: +45 23326495. Email: pb.eco@cbs.dk.

[c] Universidade Federal de Minas Gerais, Av. Antonio Carlos, 6627, 31270-901, Pampulha – Belo Horizonte, Minas Gerais, Brazil. Phone: +55 31 99212-7972. Email: analopes.ufmg@gmail.com.

Abstract:

The paper deals with benchmarking cases where highly aggregated decision making units are in the data set. It is shown that these units – consisting of sub-units which are not further known by the evaluator – are likely to receive an unjustifiable harsh evaluation, here referred to as aggregation bias. To counter this bias, we present an approach which allows to calculate the *potential sub-unit efficiency* of a decision making unit by taking into account the possible impact of its sub-units' aggregation without having disaggregated sub-unit data. Based on data envelopment analysis, the approach is operationalized in several ways. Finally, we apply our method to the benchmarking model actually used by the Brazilian Electricity Regulator to measure the cost efficiency of the Brazilian distribution system operators. For this case, our results reveal that the potential effect of the aggregation bias on the operators' efficiency scores is enormous.

Keywords:

Benchmarking; Data envelopment analysis; DEA; Aggregation bias; Potential sub-unit efficiency; Regulation

JEL:

C14; C61; C67

Conflict of interest:

Although one of the funds is supported by a company, there is no conflict of interest. The developed approach is universally applicable and of general interest for any benchmarking study.

Measuring *Potential Sub-Unit Efficiency* to Counter the Aggregation Bias in Benchmarking

1. Introduction

In efficiency analysis, e.g., in Data Envelopment Analysis (DEA) that is addressed in this paper, it is customary to work with aggregated production data. Not only non-homogenous inputs and outputs are aggregated into a few input and output categories, but also spatially distributed production entities as well as temporally consecutive production processes are aggregated, often into only one overall, consolidated unit to be evaluated. In the terminology of traditional activity analysis, this means that rows (variables) as well as columns (individual production entities/processes) are aggregated before the actual evaluation of the consolidated unit. While the latter is usually referred to as a *decision-making unit* (DMU), its individual production entities and production processes are hereinafter called *sub-units*.

The abovementioned aggregations affect the evaluation. In the context of DEA, for example, we can usually identify more inefficiency the more we aggregate the variables, since this relaxes the mathematical programs defined to reveal inefficiency. However, especially the aspect of sub-unit aggregations has not received much attention in the efficiency analysis literature. Little is known about the magnitude of the aggregation impact, and little is known about how to measure this impact. The present paper aspires to shed some light on this issue, stressing in particular the possible bias resulting from the evaluation of consolidated DMUs.

This bias occurs under the usual DEA assumption of convexity of the production possibility set. Here, a DMU can improve its overall profit by adjusting production to variations in prices over space and time, whereas its technical efficiency based on physical production data will deteriorate because a convex technology favours producing the average output using the average input. In other words, if a consolidated DMU operates several sub-units that serve different areas with different needs and characteristics, the DMU may easily appear inefficient, although it is in fact operating optimally in the different areas. Such a shortcoming, subsequently referred to as *aggregation bias*, analogously applies for different time spans of a period that are to be served differently by a DMU.

A series of plausible examples can be found which are prone to the aggregation bias, ranging

- from special area-related cases, such as the one of German savings banks whose local branches have to cover an assigned region no matter how diverse the areas of this region may be (Ahn and Le 2015),

- to the general case of companies that operate on diverse markets, e.g., worldwide with corresponding international production sites, as well as

- from special time-related cases, such as the one of Taiwanese design consultancies for which it has been shown that it is advantageous to match their the time-based business strategy with the actual conditions of the market environment (Sung, Lu and Ho 2010),

- to the general case of companies that must adjust their production to seasonal fluctuations in demand, e.g., in the tourism industry.

As a further example of where the aggregation bias matters, we will emphasise the regulatory context in the paper. On one hand, firms are regulated because they benefit from a regional monopoly. On the other hand, these firms must deal with the specifics of their particular market region. Such a region may be characterised by a significant heterogeneity of its individual service areas, e.g., due to variations of the production environment and customer structure. Consequently, a firm affected by this scenario (as a DMU) will develop different strategies for its different service areas (as sub-units) to maximise its overall profit. Due to the described aggregation bias, however, such a DMU is likely to receive an unjustifiable harsh evaluation by the regulatory authority, which does not take into account the sub-unit structure.

In addition to the negative impact of the aggregation bias on mere efficiency evaluation, at least two further issues arise:

- Aggregated evaluations are difficult to reconcile with more detailed evaluations of other evaluators of the same production system: economists using aggregate data and engineers using detailed data may obtain conflicting results; marketing analysts can be impressed by a firm's adaptability to changing market conditions, and yet, an aggregate evaluation may show deficient allocative efficiency because of the underlying convex model. This impedes communication and trust between the respective stakeholders.

- The aggregation bias induces adverse incentive effects. A contract with a DMU designed to improve technical and allocative efficiency (cf., e.g., Bogetoft and Otto 2011; Bogetoft 2012) may diminish responsiveness to variations in the market condition because adjustments to in-

puts and outputs will manifest as increased technical inefficiency in the aggregated evaluation. Hence, contracting on technical efficiency may lead to lower profitability and, more generally, poorer goal fulfilment. It is important that incentive schemes take this problem into account. For example, leading regulators of electrical networks have made corresponding amendments to their revenue cap regulation (cf., e.g., the merger approach used by the Norwegian regulator as discussed in Bogetoft 2012).

Against this background, the paper suggests an approach to counter the aggregation bias by evaluating the potential impact of sub-unit aggregations on a DMU's DEA efficiency without actually possessing disaggregated sub-unit data. Under the condition that no knowledge about the sub-units and their actual number exists, the approach involves hypothetical disaggregations to investigate whether the conclusions derived from the aggregate information could be significantly altered by more detailed information.

The impact our approach can have on efficiency analysis is verified based on data from the Brazilian Electricity Regulator used in 2015 to measure the cost efficiency of Brazilian distribution system operators (DSOs). These DSOs are confronted with a – in particular geographically – heterogeneous business environment that is susceptible to the aggregation bias. Accordingly, our findings impressively indicate the relevance the aggregation bias may have. In comparison to the results of the Brazilian DSO model, the number of DSOs identified as efficient is clearly higher when using our approach. Furthermore, the particular efficiency scores substantially increase under our approach.

The paper is organised as follows. After an overview of the related literature in Section 2, Section 3 provides a conceptual introduction to the idea of distinguishing between disaggregated sub-unit efficiency and aggregated DMU efficiency. We provide simple examples of the aggregation bias, explain how this bias arises, and outline the measure of *potential sub-unit efficiency* (*PSE*) that serves to correct a biased evaluation of DMU efficiency. As a starting point to formalising the *PSE* concept, Section 4 discusses the condition under which the aggregation bias does not arise. On this basis, Section 5 focuses on the opposite case. After proposing a general approach to measure *PSE* referring to Farrell efficiency, we elaborate an operational *PSE* measure for the case of missing sub-unit information and also simplify the resulting mixed-integer program. In Section 6, our model is applied to the case of Brazilian DSO regulation to investigate to what extent these DSOs may be affected by the aggregation bias. Final conclusions are provided in Section 7.

2. Literature Review

We are not aware of papers that are directly related to the present study. There are, however, five lines of research with links to the issues we address.

One line of studies relates to sensitivity analysis. To determine a simple, yet sufficiently detailed model, one may examine how sensitive the results (efficiency scores etc.) are to aggregations over space and time, respectively. This issue is clearly only one of several that can be addressed by a sensitivity analysis. For other applications in DEA models, see, for example, Charnes and Neralic (1990) as well as Charnes, Rousseau and Semple (1996). However, our approach does not calculate the consequences of given aggregations; rather, we seek to find the disaggregation that has the maximal effect on the results. We construct worst-case scenarios to put the evaluated DMUs in the best possible light.

Another line of related research combines the methods from efficiency analysis, including the estimation of production models using observed productions, with those reported in the production planning literature, including the use of linear programming to model networks of production processes. Based on the seminal paper of Färe and Grosskopf (2000a) about network DEA, a series of conceptual variations were provided. Broadly recognised, for example, is the contribution of Kao and Hwang (2008), who determine the efficiency of a production process as the product of sub-process efficiencies. For a recent contribution to network DEA, see, Ma and Chen (2018). Such approaches deviate from our approach by a) modelling multi-step production systems, whereas we exclusively look at parallel sub-processes, and b) assuming that data are available at the sub-process level, whereas we assume that such sub-process information is not available, thus investigating how this limitation could affect DEA efficiency scores.

A third line of inquiry is that of DEA-based merger analysis. In a series of contributions, benchmarking models were used to make ex ante predictions of the likely gains from mergers (cf., e.g., Bogetoft and Wang 2005); the approach suggested in the latter paper has been adopted by regulators in the Netherlands and Norway to guide decision-making and incentive regulation in the context of mergers in the hospital and the energy sector (cf. Bogetoft 2012). Furthermore, Andersen and Bogetoft (2007) as well as Bogetoft et al. (2007) examined the effect of allowing more general reallocations of some of the resources and services within a large number of market participants. Whereas those studies progressed from less aggregated to more aggregated production

units, the present paper goes in the opposite direction by proceeding from aggregated to dis-aggregated production units. We ask what bias may occur when DMUs are evaluated as if there were synergies, although such synergies do not in fact exist because of geographic or temporal barriers. Hence, whereas the merger literature studies hypothetical mergers, the present paper studies hypothetical disaggregations of consolidated DMUs.

Another line of research investigates the aggregation of variables (rows in traditional activity analysis formulations) and shows how this may impact the efficiency scores (e.g., Tauer 2001; Färe, Grosskopf and Zelenyuk 2004). In contrast, we investigate the aggregation of sub-units (columns in traditional activity analysis formulations). Focusing on the problem of variable ag-gregation, Färe and Zelenyuk (2002) show that aggregation may not matter, i.e., that aggregation is exact if and only if there is no allocative inefficiency in the sub-vector of aggregated inputs. Conceptually, the latter argument has similarities to our preliminary considerations in Section 4, where we assume that fully aligned sub-units lead to no additional inefficiency at the aggregate level. On this basis, we then for the first time investigate the possible impact of not fully aligned sub-units without having sub-unit data (see Section 5).

To the best of our knowledge, the most similar paper to ours is Imanirad, Cook and Zhu (2013). These authors address a problem involving several sub-processes (possibly using a subset of in-puts to produce a subset of outputs) that aggregates into the observable aggregate input usage and output production. They suggest to hypothetically split the inputs among sub-processes and to measure aggregate efficiency as a weighted average of sub-process efficiency. Their approach assumes a splitting of the total production into sub-processes, i.e., combinations of inputs that can produce combinations of outputs. The additional information about sub-processes provides a unique splitting of the outputs on different sub-processes, and the remaining issue is now how to split inputs that are used in different sub-processes. This is done endogenously to maximise the weighted efficiency of the sub-units where the weights reflect the relative share of inputs used. By contrast, our approach does not assume any knowledge of specific sub-processes transforming some input combination to a given output vector. Also, we do not allow the aggregation of inputs to be free. Rather, we assume that the sub-processes are best represented directly by the actual observations, with information about production at the aggregate level. Hence, compared to Ima-nirad, Cook and Zhu (2013), we stick to the traditional DEA philosophy of letting present data speak, i.e., assuming no information in addition to the observed (aggregate) input-output vectors.

To sum up, we add an innovative aspect to the vast DEA literature. Yet, from a broader perspective, our paper can be viewed as part of the research stream of DEA under the condition of a so-called centralised management (cf., e.g., Lozano and Villa 2004, and most recently, Afsharian, Ahn and Thanassoulis 2017). This notion covers contributions in which a central evaluator not merely puts every DMU in its best possible light but incorporates management control mechanisms into the efficiency measurement. These mechanisms incentivise the DMUs (as agents) to make decisions in accordance with the goals of the regulator (as principal). In line with this, our approach applies such a mechanism that counters the aggregation bias, preventing large companies from reorganising into smaller units only because this would lead to higher efficiency scores. The underlying idea of distinguishing between disaggregated sub-unit efficiency and aggregated DMU efficiency is illustrated in the next section, using two motivational examples.

3. Exemplary Description of the Aggregation Bias and its Quantification

Most companies comprise individual sub-units that are operated in parallel over time. When these sub-units are described in aggregate terms, the adaptation to variations in supply and demand over space and time will be suppressed. As already outlined, actually favourable sub-unit adaptions may lead to unfavourable aggregate evaluations of the organisation as a whole. In such a setting, we can distinguish between two efficiency notions:

- *Sub-unit efficiency*: Individual sub-units cannot be improved.
- *DMU efficiency*: The aggregate of the sub-units cannot be improved.

The issue we investigate is that an organisation may have fully efficient sub-units but still appears inefficient on an aggregate level – we call this the aggregation bias.

A simple example is the so-called Fox paradox (cf. Fox 1999, 2012). One version of the phenomenon is illustrated in Table 1, which describes the case of two companies, e.g., electrical network firms, both serving rural and urban customers. DMU A serves 2 rural and 4 urban customers and spends 1 on the rural and 1 on the urban customers. The unit costs (UC) of the two types of consumers are therefore $\frac{1}{2}$ and $\frac{1}{4}$, respectively. In total, DMU A has spent 2 to serve 6 customers, and the aggregate UC are $\frac{2}{6}$. The other entries can be interpreted analogously.

Table 1 Fox paradox as example for the aggregation problem

DMU	Rural UC	Urban UC	Aggregate UC
A	$^1/_2$	$^1/_4$	$^2/_6 = 0.33$
B	$^{1,5}/_2$	$^{21}/_{80}$	$^{22,5}/_{82} = 0.27$

The interesting observation follows from the fact that DMU A has

- lower rural unit costs (costs rural / rural customers) and
- lower urban unit costs (costs urban / urban customers) but
- higher overall unit costs (costs / customers).

That is, sub-unit by sub-unit, DMU A is more efficient than DMU B, and yet, at the aggregate level, DMU B is the more efficient unit. The explanation is simple: The relatively more efficient activity of serving urban customers plays a larger part in DMU B than in DMU A.

In some cases, we can consider this situation a resolvable allocation problem: DMU A allocates an excessively large share of its customers to the least efficient sub-unit. If customers are freely transferable, DMU A is indeed responsible for this misallocation, and the aggregate appraisal would be fair. The important point, however, is that in cases such as the electrical network industry (and many others, as outlined in Section 1), companies cannot be held responsible for all allocation problems. For example, they cannot freely reallocate the customers between rural and urban areas. In such a situation, the aggregate evaluation becomes biased. DMU A is blamed for performance aspects it cannot control, one of the most obvious mistakes in adequate performance evaluations. Indeed, DMU A should appear efficient because for no common composition of customers can DMU B outperform DMU A.

Färe and Grosskopf (2000b) describe how to conceptually avoid the Fox paradox by applying solely additive efficiency measures. However, this way to define efficiency is not compatible with the classical DEA approach addressed in this paper that applies a ratio efficiency measure. Hence, the aggregation bias remains a possible pitfall of DEA, as the following example illustrates.

Imagine that there are 32 DMUs, each of which has used two inputs to produce the same amount of one output. Figure 1 illustrates these DMUs as numbered, filled points, and Table 2 lists the respective data set. We observe that all DMUs except for DMU 32 are fully efficient.

Fig. 1 Simple DEA example to describe the aggregation problem
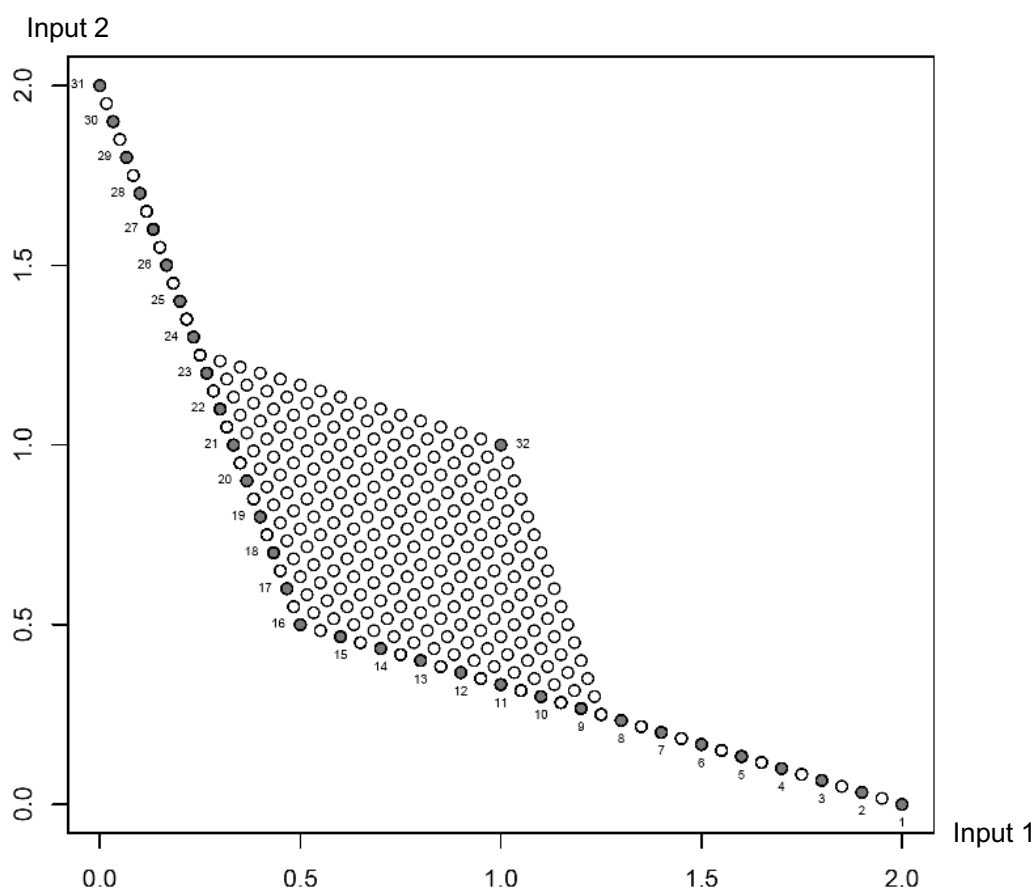
Table 2 Data set of the simple DEA example and input-oriented *PSE* scores

DMU	Input 1	Input 2	Output	*PSE$_I$** score	DMU	Input 1	Input 2	Output	*PSE$_I$** score
1	2.00	0.00	1.00	1.00	17	0.47	0.60	1.00	1.75
2	1.90	0.03	1.00	1.00	18	0.43	0.70	1.00	1.57
3	1.80	0.07	1.00	1.00	19	0.40	0.80	1.00	1.42
4	1.70	0.10	1.00	1.00	20	0.37	0.90	1.00	1.30
5	1.60	0.13	1.00	1.00	21	0.33	1.00	1.00	1.20
6	1.50	0.17	1.00	1.00	22	0.30	1.10	1.00	1.11
7	1.40	0.20	1.00	1.00	23	0.27	1.20	1.00	1.03
8	1.30	0.23	1.00	1.00	24	0.23	1.30	1.00	1.00
9	1.20	0.27	1.00	1.03	25	0.20	1.40	1.00	1.00
10	1.10	0.30	1.00	1.11	26	0.17	1.50	1.00	1.00
11	1.00	0.33	1.00	1.20	27	0.13	1.60	1.00	1.00
12	0.90	0.37	1.00	1.30	28	0.10	1.70	1.00	1.00
13	0.80	0.40	1.00	1.42	29	0.07	1.80	1.00	1.00
14	0.70	0.43	1.00	1.57	30	0.03	1.90	1.00	1.00
15	0.60	0.47	1.00	1.75	31	0.00	2.00	1.00	1.00
16	0.50	0.50	1.00	2.00	32	1.00	1.00	1.00	1.00

* PSE$_I$: input-oriented potential sub-unit efficiency

Now, assume that we also consider all consolidated versions of the efficient DMUs, i.e., all sums of two efficient DMUs including the sum of one DMU with itself. When we also allow for downscaling by $^1/_2$ to scale down the inputs of the consolidated units to give an output of 1, we obtain the other, unfilled points in Figure 1.

Let us look at one example of these consolidated units: If we add 0.5*DMU 1 and 0.5*DMU 31, we obtain the input combination (1,1), producing an output of 1. These are precisely the same inputs and outputs that characterise DMU 32. If, therefore, DMU 32 is really a consolidated DMU, it might actually be sub-unit efficient although it is inefficient at the aggregated level. In fact, if we generalise this example, we can conclude that all unfilled points are potentially sub-unit efficient.

Therewith, our conceptual basis for measuring potential sub-unit efficiency (*PSE*) can be outlined as follows:

1. A DMU may consist of sub-units, i.e., the DMU is an aggregation of its sub-units.
2. This aggregation can have the effect that the DMU's efficiency score is lower than the sub-units' efficiency.
3. We focus on such cases where this effect can be driven by an aggregation bias.
4. Assuming that the evaluator has no data about the sub-units, we propose a heuristic approach to counter the aggregation bias.
5. This heuristic, which is formally described in Section 5, provides a corrected efficiency estimation for the DMU, which we call potential sub-unit efficiency (*PSE*).
6. To calculate (different versions of) this *PSE* score, we propose to use the observed best practices in a data set.

The concept of *PSE* also takes super-efficiency into account. Concerning the numbered DMUs in our example above, e.g., the ones from 9 through 23 may represent super-efficient performance. Table 1 shows that their respective *PSE* scores are greater than 1 (calculated using the model presented in Section 5.2). Thus, these DMUs could have increased their inputs and still could have been the result of running two efficient sub-units. A *PSE* score of 1.20, for example, means that the corresponding DMU could have expanded the use of both inputs by 20% and still would have been potentially sub-unit efficient, i.e., it still could be considered an aggregation of efficient sub-units.

An interesting observation that can be derived from our approach is that centrally located DMUs have larger aggregation corrections. The reason is that they are more likely to be the result of a consolidation of sub-units with rather different input-output profiles. The more extreme DMUs have a smaller correction, i.e., when evaluated as consolidated units, their aggregation bias is smaller because they are less likely to be the result of aggregating sub-units with very different input-output profiles.

In the following section, we step-by-step formalise our *PSE* concept. As a starting point, the next section formulates and discusses the condition of sub-unit alignment under which the aggregation bias does not occur.

4. Efficiency under the Condition of Sub-Unit Alignment

Let us consider a DMU that uses m inputs $x \in \mathbb{R}_+^m \setminus \{0\}$ to produce n outputs $y \in \mathbb{R}_+^n \setminus \{0\}$. Let T be the production possibility set, with $T \in \mathbb{R}_+^{m+n}$. We assume that T is closed, convex and freely disposable. $(x,y) \in T$ is *weakly technically efficient* (or strictly non-dominated) if

$$(x - a, y + b) \notin T \text{ for all } (a, b) \in \mathbb{R}_+^{m+n}, (a, b) \gg 0 \tag{1}$$

where $(a,b) >> 0$ means that all coordinates of (a,b) are strictly positive. Let the set of weakly technically efficient productions in T be denoted $W(T)$. We note that weak efficiency is weaker than the classical economic notion of efficiency that characterises a production as efficient when no output can be increased (input can be decreased) without decreasing (increasing) another output (input). A production that can be improved in some but not all dimensions is weakly technically efficient.

We shall also say that $(x,y) \in T$ is *allocatively efficient* with respect to a price vector $(u, v) \in \mathbb{R}_+^{m+n} \setminus \{0\}$ if and only if

$$(x, y) \in \arg \max_{(x', y') \in T} vy' - ux'. \tag{2}$$

Note that allocative efficiency is often thought of in connection with given market prices of the inputs and outputs, but here we use it more generally by making allocative efficiency dependant on prices u and v, which are not determined yet.

In our theoretical analysis, it is advantageous to use the weak efficiency notion. First, the notion simplifies several of the results presented below for which we would have to assume unique solutions or solutions with strictly positive prices if we would work with ordinary efficiency to avoid picking up points on horizontal and vertical segments of the production possibility frontier (cf., e.g., Bogetoft and Pruzan 1991, Appendix A). Second, with convex sets, the weakly technically efficient productions are all those that result from optimal economic behaviour (profit maximisation) for some non-negative and non-zero price vector. Finally, in the efficiency analysis literature, it is common to work with notions of efficiency that are weaker than the classical economic one.

Referring to the last-mentioned aspect, the predominantly application of Farrell efficiency measures in DEA surely has an historical background. However, there are also practical reasons to justify their consideration. Compared to other approaches like, e.g., the slack-based measure of Tone (2001), it is easier for practitioners to interpreting them, i.e. to understanding the concept of radially measured efficiency scores. Furthermore, Farrell efficiency corresponds particularly well with findings in incentive theory. Imposing proportional changes, which is the very essence of Farrell, can under some circumstances be shown to be the optimal response of a principal who lacks information about the relative costs of different activities of his agents (see, e.g., Bogetoft 2000).

Returning to our assumptions, by disposability, $W(T)$ constitutes the "North-West" boundary points of T. Moreover, since T is closed and convex, it is the intersection of all halfspaces $\{(x',y') \mid vy - ux \leq r\}$ containing T (cf. Rockafellar 1970, Theorem 11.5). Furthermore, it is well-known that any halfspace containing T has a normal $(-u,v)$, where (u,v) is non-negative (see, e.g., Hackman 2010). From this follows the following basic observation, which is a version of the Koopmans-Gale price theorem (see, e.g., Frank 1969): *A production $(x,y) \in T$ is weakly technically efficient if and only if it is allocatively efficient with respect to some price vector $(u,v) \in \mathbb{R}_+^{m+n}\backslash\{0\}$.* Hence, the weakly technically efficient productions are allocatively efficient under some (non-negative and non-zero) price vector and vice versa.

Now, to investigate the aggregation problem, let us assume that the production of a DMU can be split into $|H|$ sub-units, $h \in H = \{1,\ldots,|H|\}$. For any h, let $(x^h,y^h) \in \mathbb{R}_+^{m+n}$ be the production vector, let T^h be the set of possible productions for h, and when applicable, let $(u^h,v^h) \in \mathbb{R}_+^{m+n}\backslash\{0\}$ be

the non-negative prices of the inputs and outputs of *h*. For simplicity, we assume that inputs and outputs are freely disposable, i.e., for any T^h we assume that $(x', y') \in T^h$, $x'' \geq x'$, $0 \leq y'' \leq y'$ $\Rightarrow (x'', y'') \in T^h$. Furthermore, we assume that T^h, $h \in H$, are closed and convex.

We can now provide a necessary and sufficient condition for aggregations to create no evaluation biases. This necessary and sufficient condition for efficiency at the aggregate level is allocative efficiency at the sub-unit level with regard to the same set of prices – we call this the alignment condition.

Theorem 1. *If T^h is convex for all $h \in H$, the aggregate production $\sum_{h \in H}(x^h, y^h)$ is weakly technical efficient at the aggregate level*

$$\sum_{h \in H}(x^h, y^h) \in W(\sum_{h \in H} T^h) \tag{3}$$

if and only if the sub-unit productions (x^h, y^h), $h \in H$, are allocatively efficient at the sub-unit level with the same rate of substitutions in all sub-units:

$$\exists (u, v) \in \mathbb{R}_+^{m+n}\backslash\{0\}: (x^h, y^h) \in \arg \max_{(x,y) \in T^h} vy - ux \ \forall h \in H. \tag{4}$$

Proof: When T^h is convex for all $h \in H$, so is $\sum_{h \in H} T^h$. Therefore, when the aggregate production is efficient, $\sum_{h \in H}(x^h, y^h) \in W(\sum_{h \in H} T^h)$, we know from the basic observation that it is allocativly efficient with respect to some price vector

$$\exists (u, v) \in \mathbb{R}_+^{m+n}\backslash\{0\}: \sum_{h \in H}(x^h, y^h) \in \arg \max_{(x,y) \in \sum_{h \in H} T^h} vy - ux \tag{5}$$

In turn, this relationship implies

$$\exists (u, v) \in \mathbb{R}_+^{m+n}\backslash\{0\}: (x^h, y^h) \in \arg \max_{(x,y) \in T^h} vy - ux \ \forall h \in H \tag{6}$$

because had we used the same price vector as in (5) and (x^h, y^h) does not solve $\max_{(x,y) \in T^h} vy - ux$ for some *h*, there would have been a better solution to (5) obtained by substituting the solution into $\max_{(x,y) \in T^h} vy - ux$ for the old (x^h, y^h) in (5). This finding shows that allocative efficiency at the sub-unit level with the same rate of substitution in all sub-units is a necessary (only if) condition for aggregate efficiency.

The sufficiency also follows from a contradiction. Let $(x^h, y^h) \in \arg max_{(x,y) \in T^h} vy - ux$, $h \in H$ be a solution to (6). Now, if $\sum_{h \in H}(x^h, y^h)$ does not solve (5), there exists an alternative

solution $\sum_{h \in H}(\tilde{x}^h, \tilde{y}^h)$ such that $v \sum_{h \in H} \tilde{y}^h - u \sum_{h \in H} \tilde{x}^h > v \sum_{h \in H} y^h - u \sum_{h \in H} x^h$. Thus, for at least one h, we have $v\tilde{q}^h - u\tilde{x}^h > vy^h - ux^h$, which contradicts the fact that we had a solution (6) to begin with. Hence, a solution to (6) yields a solution to (5), and by the basic observation, the result is weak technical efficiency. This proves the sufficiency (if) part of the theorem. □

Theorem 1 shows that to attain technical efficiency at the aggregate level, it is not enough to be technical and allocatively efficient at the sub-unit level. Optimal profit-maximising behaviour does not ensure aggregate technical efficiency. We also need for the sub-units to be aligned by a common price vector in the sense that the rates of substitution are the same in all sub-units. In more organisational terms, we can conclude that it is not sufficient to have efficient sub-units; we need goal concordance among them as well.

Note that goal concordance in this case is considered with respect to a common weighting of the inputs and outputs, which it is not necessarily advantageous. If prices differ, concordance regarding the same weights is sub-optimal. In such a case, the alignment condition in Theorem 1, which implies that scale and scope effects do not come into play, causes a discrepancy between sub-unit performance and aggregated DMU performance. In other words: in addition to optimal behaviour in the different sub-units, we need price proportionality in the sub-units to aggregate information without obscuring the evaluations.

One instance of Theorem 1 is particularly clear. If all sub-units are the same, $T^1 = T^2 = \ldots = T^{|H|}$, a necessary condition for technical efficiency at the aggregate level is that they all operate if not at the same point, then at least at the same facet, i.e., have the same supporting hyperplane. This follows immediately from the alignment condition in Theorem 1. Furthermore, if the technologies are strictly convex – at least in the relevant (efficient) part – we conclude that the aggregate production is efficient if and only if all sub-units are using exactly the same (efficient) production.

This situation illustrates how desirable adjustments to spatial/temporal variations in prices at the sub-unit level will tend to appear as technical inefficiencies at the aggregate level. Optimal economic behaviour in the sub-units leads to aggregate inefficiency if prices over space/time are not accidentally proportional. It is fair to say, therefore, that aggregate efficiency is quite unlikely and certainly not always desirable; such efficiency may come at the cost of inadequate adaptions to local variations in prices or, more generally, to variations in demand and supply.

We note that these statements regard aggregate efficiency in an absolute sense, i.e., compared to a theoretical production model. In practice, a DMU's efficiency is usually measured relative to other DMUs, and because they also may be "handicapped" by variations in the prices on the dis-aggregated level, a DMU is more likely to be (relatively) as efficient as an aggregate unit.

In the next section, we propose an approach to ensure fairer evaluations when DMUs cannot freely allocate production between different sub-units, i.e., when the condition of sub-unit alignment is not valid.

5. Potential Sub-Unit Efficiency

5.1. A *PSE* Approach Based on Farrell Efficiency

In practice, it is uncertain if inefficiency at the aggregate level is the result of slack in the sub-unit productions or the result of an aggregation of non-aligned but efficient sub-units. We may, however, ask if a DMU's observed inefficiency is sufficiently small to be explained by unobserved but potentially desirable adjustments of its sub-unit productions to local or temporal variations in prices. Taking this perspective a bit further, we may ask how much of aggregate inefficiency we can explain merely by the aggregation of sub-units.

If the observed aggregate production can be expressed as an aggregation of efficient sub-unit productions, we say that the aggregate production is characterised by *potential sub-unit efficiency* (*PSE*). More precisely, we say that $(x,y) \in T$ is *potentially* $(T^h, h \in H)$ *sub-unit efficient* (*PS* efficient) if and only if

$$\exists (x^h, y^h) \in Eff(T^h), h \in H: x \leq \sum_{h \in H} x^h, y \geq \sum_{h \in H} y^h \qquad (7)$$

where $Eff(T^h)$ is the efficient production in T^h. Hence, a *PS* efficient aggregate production of a DMU is one that results from aggregating efficient – but not necessarily aligned – sub-unit productions. To test such an aggregate production, we let the sub-units be hypothetical and construct sub-unit productions that put a DMU in the best possible light. On this basis, we can evaluate whether a production (*x*,*y*) is possibly sub-unit efficient and also measure the corresponding *PSE* score.

As always, efficiency can be operationalised in different ways. To convey our idea in a simple way, we here refer to Farrell's idea of determining input efficiency (*E*) and output efficiency (*F*) by means of proportional input and output adjustments, respectively. Based on this easy to inter-

pret concept, we can measure the *PSE* of (*x*,*y*) on the input side, *PSE_I*(*x*,*y*), as

$$PSE_I(x, y) = \max\{E | Ex \leq \sum_{h \in H} x^h, y \geq \sum_{h \in H} y^h, (x^h, y^h) \in Eff(T^h), h \in H\}. \quad (8)$$

The interpretation of this program is straightforward: We seek the largest possible expansion of the inputs used in (*x*,*y*) such that the resulting production still uses no more inputs to produce at least the same output as some combination of efficient sub-units. A score of *PSE_I*(*x*,*y*) = 1 denotes the case that (*x*,*y*) is *PS* efficient. If *PSE_I*(*x*,*y*) is greater than 1, (*x*,*y*) is in fact super-*PS* efficient. If *PSE_I*(*x*,*y*) is less than 1, even when we take limitations of an alignment between the sub-units into account, there are some savings to be made in the production (*x*,*y*).

Likewise, we can measure *PSE* on the output side by

$$PSE_O(x, y) = \min\{F | x \leq \sum_{h \in H} x^h, Fy \geq \sum_{h \in H} y^h, (x^h, y^h) \in Eff(T^h), h \in H\}. \quad (9)$$

The interpretation of this program is again simple: We seek the largest possible contraction of the outputs produced in (*x*,*y*) such that the resulting production still uses no more inputs to produce at least the same output as some combination of efficient sub-units. A score of *PSE_O*(*x*,*y*) = 1 denotes the case that (*x*,*y*) is *PS* efficient. If *PSE_O*(*x*,*y*) is less than 1, (*x*,*y*) is super-*PS* efficient because there exists a combination of efficient sub-units that produce less than *y* using at least the same inputs *x*. If *PSE_O*(*x*,*y*) is greater than 1, there is some potential to expand the outputs in (*x*,*y*), even when we allow for limitations of an alignment between the sub-units.

5.2. Operational PSE Measures in Case of Missing Sub-Unit Information

In many evaluation settings, detailed information about the sub-units is not available. We do not know the set of sub-units *H* or the production possibilities in these units, T^h. We shall now discuss how to measure *PSE* when such sub-unit information is missing.

Taking a usual benchmarking study as a starting point, we assume that each of *K* DMUs, DMU^1,...,DMU^K have used inputs $x^k \in \mathbb{R}_+^m$ to produce outputs $y^k \in \mathbb{R}_+^n$. Let *T* be the underlying production possibility set, and let *T** be an estimate of *T* based on $\{(x^k, y^k) | k \in K\}$. If, for example, we assume convexity, free disposability and weakly increasing (non-decreasing) returns to scale (i.e., $(x, y) \in T$ and $k \geq 1$ implies $(kx, ky) \in T$), the minimal extrapolation estimate of *T* using DEA would be

$$T^* = \{(x, y) \in \mathbb{R}_+^{m+n} | x \geq \sum_{k \in K} \lambda^k x^k, y \leq \sum_{k \in K} \lambda^k y^k, \sum_{k \in K} \lambda^k \geq 1, \lambda^k \in \mathbb{R}_0\} \quad (10)$$

To cope with missing information about the sub-unit production possibilities T^h, we propose using the basic idea of non-parametric benchmarking, namely relying on observed best practices. One simple approach is to project the original observations on the efficient frontier of T^* and to only use combinations of these observations in the evaluation of *PSE*. To this end, let the Farrell input efficiency of DMU^k be

$$E^k = \min\{E \,|\, (Ex^k, y^k) \in T^*\} \tag{11}$$

and consider

$$Eff(T^h) = \{(E^k x^k, y^k),\, k \in K\} \; \forall h \in H. \tag{12}$$

That is, we let the efficient outcomes of sub-unit h be the efficient versions of the original DMU observations. Thereby, an observation (x^k, y^k) is *PS* efficient if it is the sum of efficient versions of the actual observations. This *direct aggregation approach* leads to a *PSE* measure, which is based on a minimum of speculations regarding what is feasible and is easy to interpret. The greater $Eff(T^k)$, the more *PS* efficient the DMUs will appear, i.e., we will obtain larger PSE_I and smaller PSE_O scores. We can therefore say that the approach of relying solely on the efficient versions of the original observation is cautious in the sense that it does not lead to excessively good *PSE* evaluations.

Using the described approximation of the efficient sub-unit outcomes, the calculation of PSE_I can be outlined in four steps:

Step 1: Compute the Farrell input efficiencies E^k of each of the original observations (x^k, y^k), $k \in K$.

Step 2: Determine the efficient sub-unit outcomes as

$$Eff(T^h) = \{(E^k x^k, y^k),\, k \in K\} \; \forall h \in H. \tag{13}$$

Step 3: Determine the *PSE* reference technology as all possible additions of efficient observations

$$\tilde{T} = \{(\textstyle\sum_{h \in H} \tilde{x}^h, \sum_{h \in H} \tilde{y}^h) \,|\, (\tilde{x}^h, \tilde{y}^h) \in Eff(T^h) \; \forall h \in H\}. \tag{14}$$

Step 4: Determine the *PS* input efficiency of the original observations as

$$PSE_I(x, y) = \max\{E \,|\, Ex \le \bar{x}, y \ge \bar{y}, (\bar{x}, \bar{y}) \in \tilde{T}\}. \tag{15}$$

This formulation of the direct aggregation approach can be simplified. The last three steps essentially correspond to the solution of the mixed-integer program

$$PSE_I(x,y) = \max_{E,\lambda^1,\dots,\lambda^K} E$$

$$s.t. \quad Ex_i \leq \sum_{k=1}^{K} \lambda^k E^k x_i^k, \quad i = 1,\dots,m$$

$$y_j \geq \sum_{k=1}^{K} \lambda^k y_j^k, \quad\quad j = 1,\dots,n$$

$$\lambda^k \in \{0,1,2,3,\dots\}, \quad\quad k,\dots,K$$

$$\sum_{k=1}^{K} \lambda^k \leq |H|. \quad\quad\quad (16)$$

In this program, we look for a combination of efficient productions ($E^k x^k, y^k$), $k = 1,\dots, K$ such that the combination uses more inputs to produce less outputs than (Ex,y).

Of course, the procedure outlined above can also be executed using projections in the output direction in Steps 2 and 4:

Step 2*: Determine the efficient sub-unit outcomes as

$$Eff(T^h) = \{(x^k, F^k y^k), k \in K\} \; \forall h \in H. \quad\quad (17)$$

Step 4*: Determine the *PS* output efficiency of the original observations as

$$PSE_o(x,y) = \min\{F | x \leq \bar{x}, Fy \geq \bar{y}, (\bar{x}, \bar{y}) \in \tilde{T}\}. \quad\quad (18)$$

The mixed-integer program solving PSE_O becomes

$$PSE_o(x,y) = \min_{F,\lambda^1,\dots,\lambda^K} F$$

$$s.t. \quad x_i \leq \sum_{k=1}^{K} \lambda^k F^k x_i^k, \quad i = 1,\dots,m$$

$$Fy_j \geq \sum_{k=1}^{K} \lambda^k y_j^k, \quad j = 1,\dots,n$$

$$\lambda^k \in \{0,1,2,3,\dots\}, \quad k,\dots,K$$

$$\sum_{k=1}^{K} \lambda^k \leq |H|. \quad\quad\quad (19)$$

In applications, the number of underlying sub-units H to consider may be uncertain. Then, we have different options. One possibility is to choose the most restrictive option and only allow DMUs to be hypothetically decomposed into two sub-units, i.e., $H = 2$. This option leads to cautious results in the sense that it yields the smallest set of *PS* efficient outcomes and therefore the smallest scores of the *PSE_I* measure and the largest scores of the *PSE_O* measure. The DEA example in Section 3 was explained using this version of the direct aggregation approach.

Another possibility is to choose the most flexible option and say that we allow DMUs to be hypo-thetically decomposed into any number of sub-units, i.e., H = any natural number. This option yields the largest set of *PS* efficient outcomes and therefore the largest scores of the PSE_I meas-ure and the smallest scores of the PSE_O measure.

5.3. A Simplified (relaxed) Approach

The aggregation approach is conceptually simple because it directly constructs possible *PS* effi-cient outcomes by adding together efficient versions of actual outcomes. Still, computationally, this approach may seem complicated because it involves mixed-integer programming. It may therefore be interesting to study the following simplified version of the mixed-integer program:

$$PSE_I(x, y) = \max_{E, \lambda^1, \dots, \lambda^K} E$$

$$s.t. \quad Ex_i \leq \sum_{k=1}^{K} \lambda^k E^k x_i^k, \quad i = 1, \dots, m$$

$$y_j \geq \sum_{k=1}^{K} \lambda^k y_j^k, \qquad j = 1, \dots, n$$

$$\lambda^k \in \mathbb{R}_0, \qquad\qquad k, \dots, K$$

$$\sum_{k=1}^{K} \lambda^k \geq 1. \tag{20}$$

This problem is relaxed by allowing the λ values to be real numbers as opposed to integers and by removing the upper constraint on the sum of these values.

The program is again straightforward to interpret. It involves performing an output-oriented DEA efficiency analysis of (*x,y*) in which we assume convexity, increasing (non-decreasing) returns to scale (IRS), and free disposability. Furthermore, the inputs and outputs are reversed, i.e., the in-puts x are treated as outputs and outputs y are treated as inputs.

Because the relaxation leads to slightly larger scores of PSE_I, the efficiency measures calculated in this manner may exceed the scores calculated using the direct aggregation approach. Hence, the DMUs will tend to appear more efficient. As we will observe, however, the relaxation may be modest. In fact, in the example presented in Section 6, the relaxed problem leads to the same scores yielded by the direct aggregation approach with $H = 2$.

Instead of considering the relaxed formulation as an approximation, it is possible to motivate the formulation in its own right. One approach could be to assume that the underlying technology is an IRS technology. The weighted sum of efficient sub-units can now be rewritten as follows:

$$(\sum_{k=1}^{K} \lambda^k E^k x^k, \sum_{k=1}^{K} \lambda^k y^k) = (\sum_{k=1}^{K} \alpha^k r^k E^k x^k, \sum_{k=1}^{K} \alpha^k r^k y^k) \tag{21}$$

where $r^k = \text{ceiling}(\lambda^k) \in \{1,2,3,...\}$ and $0 \leq \alpha^k = \lambda^k/r^k \leq 1$ for all $k=1,…,K$. Here, the ceiling function ceiling(z) is the smallest integer not less than z. Hence, the reference unit used to evaluate the *PSE* can be interpreted as the result of two operations:

- *Downscaling:* The efficient versions of the original observations can be downscaled, making them possibly super-efficient by the increasing return to scale assumption.

- *Aggregation:* The reference unit can be any direct aggregation of a finite number of efficient and possibly super-efficient sub-units.

Hence, if we accept the IRS assumption (as in the following example), the simplified approach is conceptually easy to motivate on its own.

6. Application to the Brazilian DSO Model

6.1. Motivation

The Brazilian distribution system operators (DSOs) are regulated on the basis of a DEA model with weight restrictions to determine efficient cost levels (cf. ANEEL 2015). This example not only serves to illustrate our approach based on real-world data but also sheds some light on actual issues of benchmarking in the Brazilian energy distribution sector.

First, the mere size of the Brazilian DSOs entails a heterogeneous business environment for delivering their services. In particular, the DSOs benchmarked by the Brazilian regulator can be found in areas that range from quite flat to very hilly, from really dry to extremely humid and from landscapes with sparse vegetation to those covered by woods. Facing these considerably different geographical conditions, it is likely that many of the DSOs should in fact be regarded as consolidations of diverse sub-DSOs that have limited possibilities to create synergies. If this is the case, the evaluations based on the Brazilian DSO model may be affected by the aggregation bias.

Second, the fact that weight restrictions are used in the Brazilian DSO model may mitigate the heterogeneity problem because the resulting isoquants attain a lower curvature (this can be intuitively seen by looking at Figure 1: the more linear the isoquants, the smaller the set of *PSE* points). Therefore, although the aggregation bias is a reasonable presumption, its importance can only be evaluated by a numeric analysis based on our new approach.

6.2. The Brazilian DSO Model

The Brazilian DSO regulation is in many ways in line with the international literature on regulatory benchmarking. Corresponding models typically apply a series of indicators of the capacity provided, the transport work undertaken and the customer services delivered as cost drivers. The respective input and outputs used in the Brazilian DSO model are shown in Table 3, which also indicates the tasks covered by the different cost drivers and provides a basic explanation of them.

Table 3 Brazilian DSO model variables

Model variables	Covered task	Variable explanation
Input	**Saving of …**	
x_OPEX_adjusted	OPEX = operational expenditure	Sum of expenses, including personal, materials, specific taxes and insurances, outsourced services as well as other expenses, adjusted by the regional salary level
Output	**Provision of …**	
y_Underground_all_tension_levels	Physical assets	Total length of underground electricity distribution lines, irrespective of their voltage level
y_Air_distribution_network	Physical assets	Total length of overhead electricity distribution lines with low voltage level
y_High_network	Physical assets	Total length of overhead electricity distribution lines with high voltage level
y_Averaged_market	Transport service	Sum of MWh provided, weighted by the respective share of controllable costs
y_Consumers_number	Customer service	Number of consumers served
z_Neg_non_technical_losses_adjusted	Quality	Max (losses due to theft or fraud – respective regulatory target; 0) · low tension supply
z_Neg_interruption_adjusted	Quality	Max (average interruption duration – respective average benchmark target; 0) · number of customers

The use of physical assets to capture capacity provision is quite common in regulatory practice; although these assets are rather a means to provide the ultimate services, they can serve as reliable cost drivers, since it is unlikely that they are considerably manipulated "to play the regulation". It is also noteworthy that the model does not contain direct information about the characteristics of service areas, such as precipitation and vegetation, although these conditions vary considerably from DSO to DSO as well as across the areas serviced by the individual DSOs.

Another remarkable feature of the Brazilian DSO model is the inclusion of quality indicators as negative outputs. This direct consideration of quality differs from the usual approaches that handle quality issues indirectly by second-stage corrections or add-on regulatory instruments. It is worth mentioning that ANEEL explicitly refers to the quality indicators as positive non-controllable inputs, but just models them as negative outputs (ANEEL 2015, p. 12 and 48). Mathematically, this is the same (Bogetoft and Otto 2011, pp. 119-120). Hence, the DEA model applied by ANEEL (and therefore adopted by us) with negative outputs provides the same results as would have been provided by the respective model with non-controllable inputs. Even though the latter would include three inputs, we would still measure cost efficiency, as still only the cost input would be reduced by the respective input-oriented Farrell model.

The Brazilian DSO model also differs from common regulatory benchmarking models by the use of restrictions on the dual weights of the respective DEA problem. In total, seven such restrictions are used, as shown in Table 4. The two restrictions A and C limit the possible rate of substitution between outputs, whereas the remaining five restrict the output costs for individual outputs compared to the input OPEX (operational expenditure). The first two constraints are so-called Type I assurance regions, whereas the latter five are Type II assurance regions.

Table 4 Weight restrictions used in the Brazilian DSO model

| Restriction	Lower limit		Ratio		Upper limit
A	1	<	y_Underground_all_tension_levels/y_Air_distribution_network	<	2
B	0.58	<	y_Air_distribution_network/x_OPEX_adjusted	<	2.2
C	0.4	<	y_High_network/y_Air_distribution_network	<	1
D	0.001	<	y_Averaged_market/x_OPEX_adjusted	<	0.06
E	0.03	<	y_Consumers_number/x_OPEX_adjusted	<	0.145
F	0.01	<	z_Neg_Non_technical_losses_adjusted/x_OPEX_adjusted	<	0.15
G	0	<	z_Neg_interruption_adjusted/x_OPEX_adjusted	<	0.002

Weight restrictions can be considered either as an expression of preferences or as an expression of partial information about rates of substitutions. For example, the last restriction listed in Table 4 can be an expression of the fact that the value of avoiding an hour of electricity loss cannot exceed 0.002 kBRL, i.e., that the value of an hour of lost electricity cannot exceed 2 BRL. Alternatively, the restrictions can be an expression that the actual costs of cutting down on the hours of interruption is never higher than 2 BRL per hour. Note that it is not known whether the re-

strictions are actually expressions of regulatory preferences or of specific knowledge of cost effects (Bogetoft and Lopes 2015).

6.3. Findings

The use of weight restrictions is interesting with respect to the aggregation bias because these restrictions lead to more linear isoquants, which one would expect to limit the bias. In that respect, it can be determined that the constraints have a non-trivial impact on the Brazilian DSO model results, i.e., the constraints actually matter. For the 61 DSOs of our data set, this is illustrated in Figure 2. Here, the model results obtained using weight restrictions (the monotonically increasing black points) are compared with the pure IRS scores obtained without weight restrictions (the upper series of grey points).

Fig. 2 Impact of weight restrictions in the Brazilian DSO model
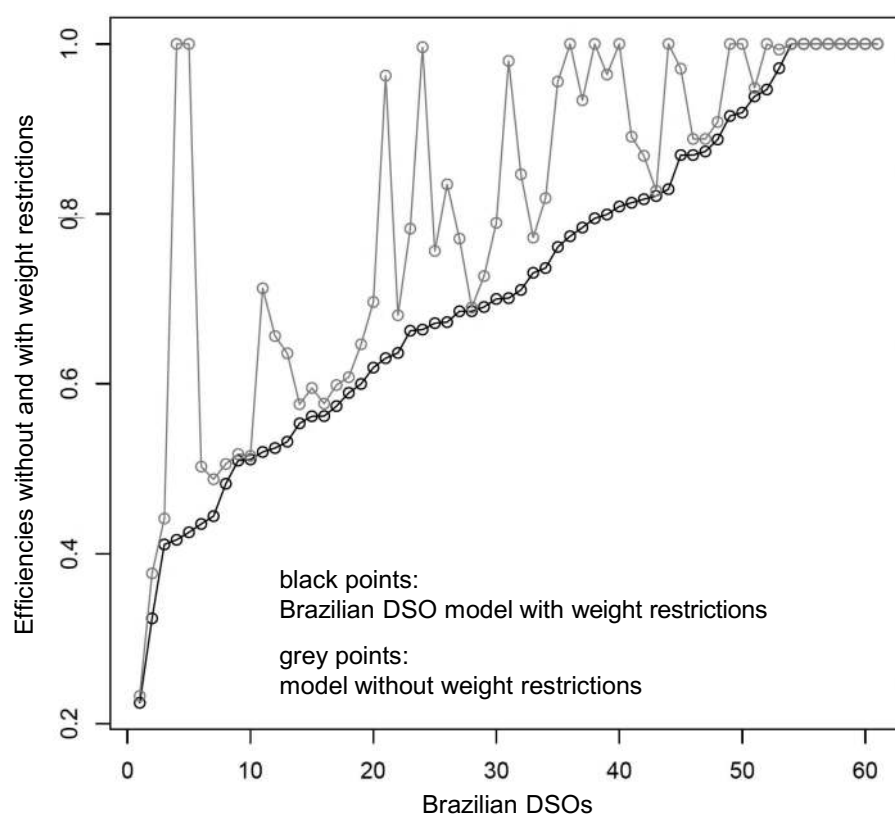
Next, we have calculated the PSE_I scores of the Brazilian DSOs using our simplified (relaxed) approach. The results are shown in Figure 3. Here, the DSOs are sorted from the smallest to largest PSE_I score. As explained above, such a PSE value quantifies the increase in costs (i.e., the

expansion factor we can multiply on actual costs) that is possible on condition that the respective DMU remains sub-unit efficient. We observe that a large share of the DSOs can in fact be considered as *PS* efficient. Only 13 of the 61 DSOs remain inefficient, with a PSE_I score less than 1, which means that 48 DSOs are classified as fully *PS* efficient (and many of them are super-efficient). By means of the Brazilian DSO model, only 8 DSOs were classified as efficient. We also observe that nearly half of the DSOs have PSE_I scores greater than 1.5, suggesting that they could in fact have increased their OPEX by 50% and that it would still be possible to consider them as sub-unit efficient.

Fig. 3 PSE_I scores
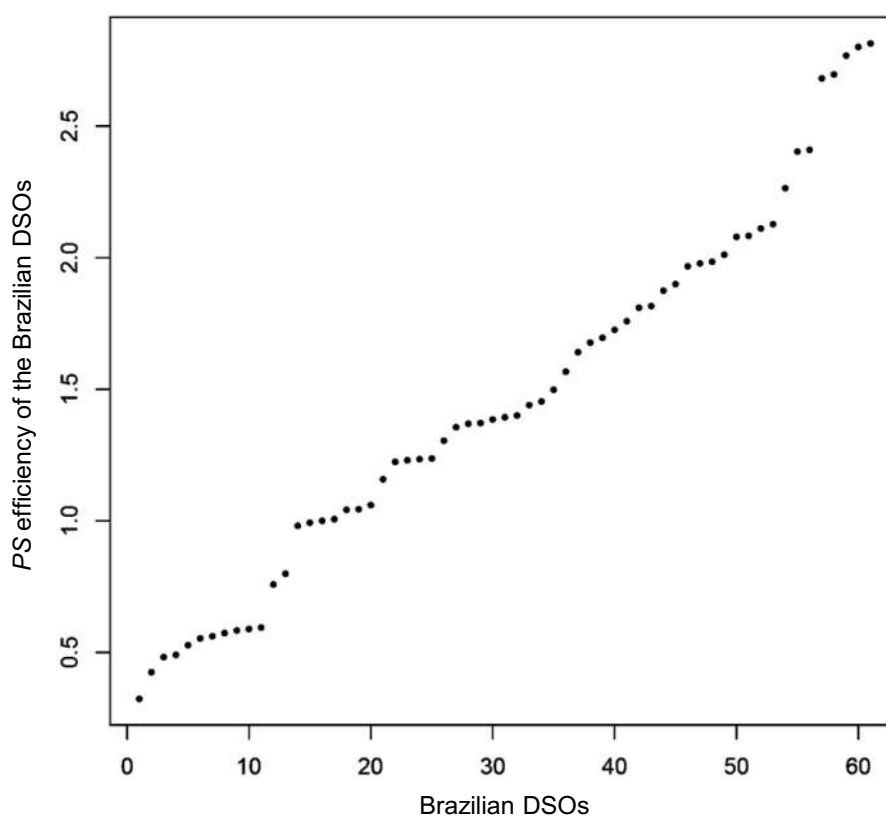
The effects of applying our approach are dramatic. Most DSOs obtain significantly better scores when we consider them as consolidated units and investigate whether they could in fact be decomposed into fully efficient sub-units. This finding is illustrated in Figure 4, in which we compare the Brazilian DSO model efficiencies (the monotonic series of black points) with the PSE_I efficiencies derived from the simplified approach (the upper series of non-monotonic grey points).

Fig. 4 Comparison of efficiency scores: the Brazilian DSO model versus the *PSE* model
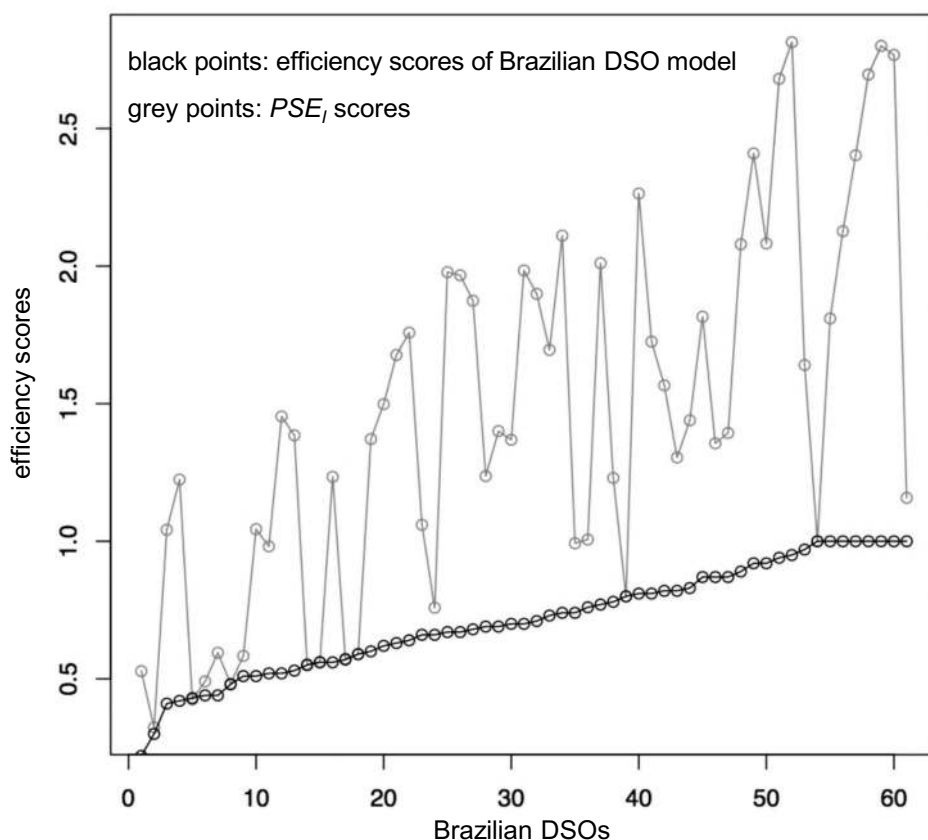
It is obvious that the derived *PS* efficiency scores are very lenient on the DSOs. This suggests that one should consider restricting the number of sub-units *H* in which the *PSE* analysis is allowed to hypothetically disaggregate the DSOs. However, our goal was to demonstrate that the aggregation of data at the DSO level can have a huge impact on the results, i.e., that the potential aggregation bias can be enormous.

7. Conclusions

In this paper, we have argued that the presence of highly aggregated organisational units in a benchmarking study may skew the results. Such DMUs are likely to receive excessively harsh evaluations. We have illustrated this aggregation bias and reflected upon the condition under which the bias does not occur, namely the alignment condition. Only with aligned productions of a DMU's sub-units, an aggregation of these productions does not affect the efficiency analysis of the DMU. Basically, price proportionality with respect to the sub-units is needed to allow for an exact aggregation of their productions without obscuring the evaluation at the DMU level.

As explained in the paper, this condition may be problematic in many real-world cases where DMUs have to manage sub-units with different business environments resulting from, e.g., different locations or periods. For such cases, we propose a DEA-based approach for compensating for the possible aggregation bias by calculating a DMU's *potential sub-unit efficiency – PSE*. This concept allows to measuring the extent to which the respective DMU can be viewed as an aggregation of efficient sub-units. If an input-oriented *PSE* score, on which we focused on, has a value less than one, it indicates that activities are not performed efficiently, even accounting for given limitations of an alignment between the sub-units. To address this effect, we elaborated how to determine *PSE* scores under different assumptions.

As an example, we applied the *PSE* concept to the DEA model used by the Brazilian Electricity Regulator in 2015 to measure the cost efficiency of the Brazilian distribution system operators (DSOs). Because of the size of these DSOs and the heterogeneity of their service areas, it is highly likely that many of the DSOs are in fact subject to biased evaluations. Our numerical results showed that the biases may be considerable. In comparison to the results of the Brazilian DSO model, the number of DSOs classified as efficient significantly increased, along with a substantial increase in many of the efficiency scores.

The implications of our findings are twofold. From the perspective of a central evaluator, e.g., a regulator, it is important to be aware of a possible aggregation bias. It seems necessary to investigate whether there are good reasons that the DMUs to be analysed operate sub-units in different business environments that require different strategies for performing optimally. In this case, incorporating the *PSE* concept into the particular efficiency analysis is a helpful control mechanism to address that issue, resulting in fairer and broader accepted evaluations.

Our findings can also be of great relevance to particular DMUs under evaluation, since the impact of the aggregation bias on efficiency scores was shown to be potentially enormous. On the one hand, it might be in the interest of affected companies to prove that a benchmarking analysis without addressing the bias would be flawed. On the other hand, companies may also react strategically, since our findings imply that 'playing the regulation' by reorganising into smaller sub-units may have a considerable payoff.

As a possibility for further research, our findings could be associated with bootstrapping in DEA (cf. the seminal paper of Simar and Wilson 1999). We speculate that uncertainty, as estimated by

bootstrapping, is largely inversely related to the extent of the consolidation bias. Although we outlined in Section 3 that the aggregation bias tends to increase from more to less extreme types of DMUs, the bias correction that can be derived from a bootstrapping exercise has the opposite tendency. This finding indicates that DMUs that involve more uncertainty in a typical efficiency analysis are less likely to have a large aggregation bias. Vice versa, DMUs that involve less uncertainty in the technical evaluations are more likely to have a large aggregation bias. A thorough investigation of this topic might be fruitful.

References

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

Ahn H, Le MH (2015) DEA efficiency of German savings banks: evidence from a goal-oriented perspective. J Bus Econ 85:953–975

Andersen J, Bogetoft P (2007) Gains from quota trade: theoretical models and an application to the Danish fishery. Eur Rev Agric Econ 34:105–127

ANEEL (Agẽncia Nacional De Energia Elétrica) (2015) Metodologia de custos operacionais [operational costs methodology]. Technical note 66/2015. Brasilia

Bogetoft P (2012) Performance benchmarking: measuring and managing performance. Springer, New York

Bogetoft P (2000) DEA and activity planning under asymmetric information. J Prod Anal 13:7–48

Bogetoft P, Boye K, Neergaard-Petersen H, Nielsen K (2007) Reallocating sugar beet contracts: can sugar production survive in Denmark? Eur Rev Agric Econ 34:1–20.

Bogetoft P, Lopes A (2015) Comments on the Brazilian benchmarking model for energy distribution regulation: fourth cycle of tariff review – technical note 407/2014. nespufmg.com.br/content/upload/p/d3d9446802a44259755d38e6d163e820.pdf. Accessed 16 December 2017

Bogetoft P, Otto L (2011) Benchmarking with DEA, SFA, and R. Springer, New York

Bogetoft P, Pruzan, P (1991) Planning with multiple criteria. North-Holland, Amsterdam

Bogetoft P, Wang D (2005) Estimating the potential gains from mergers. J Prod Anal 23:145–171

Charnes A, Rousseau J, Semple J (1996) Sensitivity and stability of efficiency classifications in Data Envelopment Analysis. J Prod Anal 7:5–18

Charnes A, Neralic, L (1990) Sensitivity analysis of the additive model in Data Envelopment Analysis. Eur J Oper Res 48:332–341

Färe R, Grosskopf S (2000a) Network DEA. Socio Econ Plan Sci 34:35–49

Färe R, Grosskopf S (2000b) Outfoxing a paradox. Econ Lett 69:159–163

Färe R, Grosskopf S, Zelenyuk V (2004) Aggregation bias and its bounds in measuring technical efficiency. Appl Econ Lett 11:657–660

Färe R, Zelenyuk V (2002) Input aggregation and technical efficiency. Appl Econ Lett 9:635–636

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

Fox KJ (1999) Efficiency at different levels of aggregation: public vs. private sector firms. Econ Lett 65:173–176

Fox KJ (2012) Problems with (dis)aggregating productivity, and another productivity paradox. Ann Oper Res 37:249–259

Frank CR Jr (1969) A generalization of the Koopmans-Gale theorem on pricing and efficiency. Int Econ Rev 10:488–491

Hackman ST (2010) Production economics: integrating the microeconomic and engineering perspectives. Springer, Berlin/Heidelberg.

Imanirad R, Cook WD, Zhu J (2013) Partial input to output impacts in DEA: production considerations and resource sharing among business subunits. Nav Res Logist 60:190–207

Kao C, Hwang S-N (2008) Efficiency decomposition in two-stage Data Envelopment Analysis: an application to non-life insurance companies in Taiwan. Eur J Oper Res 185:418–429

Lozano S, Villa G (2004) Centralized resource allocation using Data Envelopment Analysis. J Prod Anal 22:143–161

Ma, J, Chen L (2018) Evaluating operation and coordination efficiencies of parallel-series two-

stage-system: a data envelopment analysis approach. Exp Sys Appl 91:1–11

Rockafellar RT (1970) Convex analysis. Princeton University Press, Princeton

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

Sung T-J, Lu Y-T, Ho S-S (2010) Time-based strategy and business performance under environmental uncertainty: an empirical study of design firms in Taiwan. Int J Des 4(3):29–42

Tauer LW (2001) Input aggregation and computed technical efficiency. Appl Econ Lett 8:295–297

Tone K (2001) A slacks-based measure of efficiency in data envelopment analysis. Eur J Oper Res 130:498–509