DEPARTMENT OF FOOD AND RESOURCE ECONOMICS
UNIVERSITY OF COPENHAGEN

IFRO Working Paper

Implications of Aggregation Uncertainty in DEA

Emil Heesche
Mette Asmild

2022 / 02

IFRO Working Paper 2022 / 02

Implications of Aggregation Uncertainty in DEA

Authors: Emil Heesche, Mette Asmild

JEL-classification: C02; C14; C51; C52; C61; C67; L51

Published: March 2022

See the full series IFRO Working Paper here:
www.ifro.ku.dk/english/publications/ifro_series/working_papers/

Department of Food and Resource Economics (IFRO)
University of Copenhagen
Rolighedsvej 23
DK 1958 Frederiksberg DENMARK
www.ifro.ku.dk/english/

Implications of Aggregation Uncertainty in DEA

Emil Heesche[1], Mette Asmild[2]

Abstract

Researchers and practitioners who use Data Envelopment Analysis often want to incorporate several inputs and outputs in their model to consider as much relevant information as possible. However, too many inputs and outputs can result in the well-known dimensionality problem referred to as the "curse of dimensionality". Several studies suggest how to solve, or at least reduce, this problem. One solution is to aggregate the inputs and outputs before using them in the model.

This paper examines the implications when the methods used to aggregate the inputs and outputs contain uncertainty. The uncertainty can, for example, be price uncertainty if we use input and/or output prices for the aggregation.

We show that the implications for a unit under analysis depend entirely on its input and output mixes relative to those of its peers, and that the implications are higher the more heterogeneous the sector is. As an example, we use the Danish benchmarking regulation of the waste water companies. We find that uncertainty in the regulator's aggregation scheme does not, on average, influence the companies' efficiency scores a lot. Still, individual companies can be greatly affected by this uncertainty.

Keywords: Data Envelopment Analysis; Regulation; Aggregation Uncertainty; Permutation Tests

[1] University of Copenhagen, Department of Food and Resource Economics, Rolighedsvej 25, 1958 Frederiksberg C Denmark. ehe@kfst.dk (corresponding author)

[2] University of Copenhagen, Department of Food and Resource Economics, Rolighedsvej 25, 1958 Frederiksberg C Denmark. meas@ifro.ku.dk

1 Introduction

Utility providers are often subject to economic regulation because they are local monopolies. The regulation is intended to reduce consumer prices and increase quality by imitating the economic incentives found in more competitive markets.

Several different regulation schemes exist, many of which (at least in Europe) are based on the so-called revenue cap regulation. In this type of regulation, the regulator decides the companies' maximum allowed revenue, and it is hereafter up to the companies to reduce their costs accordingly to maximize their profit.

To set a revenue cap, the regulators can, for example, use benchmarking models such as Data Envelopment Analysis (Charnes, Cooper, & Rhodes, 1978) and Stochastic Frontier Analysis (Meeusen & Broeck (1977), Aigner, Lovell, & Schmidt (1977)). These models are used to force the most inefficient companies to catch up with the more efficient companies by setting the revenue caps such that they are equal[3] to the efficient companies' costs. In that way, the regulator introduces pseudo competition, which gives the companies incentives to reduce their costs, to maximize their profit. In more advanced regulatory schemes, the benchmarking model can also account for the quality of the product by allowing a higher revenue cap if the quality is high and vice versa. This paper focuses on the Data Envelopment Analysis (DEA) approach as currently used in the regulation of the Danish waste water companies.

The importance of benchmarking in utility regulations has been discussed in several papers, e.g. Agrell & Bogetoft (2018), Agrell & Bogetoft (2017), Banker, Førsund, & Zhang (2017), Bjørner & Jakobsen (2021), Goh & See (2021), Heesche & Asmild (2020), Heesche & Bogetoft (2021) and Thanassoulis (2000).

In Data Envelopment Analysis, it is standard practice to aggregate input and output measures due to the so-called "curse of dimensionality" (Kneip, Simar, & Wilson (2016)), Dyson, et al. (2001), Simar & Zelenyuk (2018)). While these aggregations can solve the dimensionality problem, they have important implications. These implications often depend on the specific context, how the aggregations are done, and the purpose of the analysis. In this paper, we examine some of the more general implications and look at the empirical implications in the model used in the Danish waste water regulation.

Zelenyuk (2020) describe three methods to aggregate inputs and outputs: The index number approach, the correlation-based approach and the price-based approach.

The index number approach covers many different methods spanning from axiomatic statistical techniques to more economic theories. For example, one can use a productivity index to proxy the aggregation function. See Zelenyuk (2020) and the references therein for a detailed discussion.

In the correlation-based approach, the aggregations are based on the correlations between the variables. The intuition is that if two or more variables are highly correlated, only one of these is needed. More formally, Principal Component Analysis (PCA) can be used to transform the correlated variables into new variables, based on their correlations, to minimize the number of needed variables without losing too much information. This method requires that the variables are highly correlated and that one will only aggregate a few of them – as the number of variables to be aggregated increases, the information loss will also

---

[3] In reality, the revenue caps are often more complicated and is not necessarily equal to the efficient companies costs but is still based on the costs.

increase. The information loss will be lower for highly correlated variables but still increasing. Another problem with the correlation-based approach is that it is hard to interpret the results because the companies cannot identify their raw data after it is transformed into principal components. They can, therefore, not as quickly understand the comparisons between themselves and their peers nor interpret the dual multipliers.

Lastly, the price-based approach aggregates input variables by using input prices and output variables using output prices. The aggregated inputs measure the companies' total costs, and the aggregated outputs measure the total revenue.[4] Zelenyuk (2020) shows that a DEA model using these aggregations calculate the companies' total inefficiency, which is defined as the sum of the technical inefficiency and allocative efficiency. In this paper, we focus on the price-based approach.

In the regulation of the Danish waste water companies, the regulator (DWRA), amongst other things, uses a DEA model to benchmark the companies. They use the companies' total controllable cost (hereafter costs) as input and the so-called OPEX and CAPEX net volumes as outputs. The net volumes are aggregations of several cost drivers. This can, for example, be the number of customers, the length and size of their pipes or the volume of waste water the companies treat. DWRA argues that there are too many cost drivers to be handled as individual outputs, so they should be aggregated. The OPEX net volume aggregates the operational cost drivers, and the CAPEX net volume aggregates the cost drivers for the companies' capacity. DWRA uses 26 OPEX cost drivers and 380 CAPEX cost drivers.

Using the companies' total controllable cost as a single input instead of considering multiple inputs, follows the price-based approach for input aggregation. This means that DWRA calculates the companies' total input inefficiency. It can be argued that this is preferable in the waste water regulation, rather than only calculating the technical inefficiency, because the goal of the regulation is to minimize the waste water prices through lower industry costs. It is, therefore, important that the companies focus on reducing their costs concerning both the technical and allocative input efficiency (the total inefficiency).[5]

However, on the output side, it is not necessarily meaningful to use the price-based approach. If all the outputs are fixed, the companies cannot influence their allocative efficiency, so this should not enter into the regulation. DWRA should, therefore, not use the price-based approach to aggregate the fixed outputs. Another reason is a political desire for the waste water companies not to gain any profit. In fact, all profit needs to be paid back to the consumers over time. Therefore, the companies should not have any incentives based on output prices that represent revenue. If, instead, the output prices represented the society's value of the outputs, it could be relevant for the regulator to use that information.

In addition, the output prices in the waste water sector are unknown. For example, there is no market for installed waste water pipes; the companies do not sell their pipes, but they make the pipes available for the consumers, and indirectly incorporate this into the waste water prices. Otherwise, the consumers would have to buy the pipes first and hereafter buy the right to have a quantity of waste water running through

[4] This requires, of course, that all inputs (outputs) are taken into consideration.

[5] Note that Zelenyuk (2020) uses "the law of one pricing" to argue that it is acceptable to use a standard input price across all companies. However, in the Danish waste water sector, the regulator uses the companies' realized costs and does, therefore, not need any assumption of "the law of one pricing".

the pipes. Because the output prices are unknown, it is not be possible to use these, even if it is advantageous.

Therefore, DWRA uses another method to aggregate the cost drivers, which is based on the price approach, but without the complications above. DWRA aggregate the cost drivers by using standardized input prices instead of output prices. The model interpretation is, thus, how much the companies actually spend compared to the expected costs. We describe and use DWRA's method in section 5. Until then, the details of how the aggregation prices are calculated are not relevant as we for now examine the price approach more generally.

This paper examines the implications of aggregating output variables using the price-based approach. However, most of our results are also valid for any other aggregation scheme. First, we illustrate how the technology set changes with the aggregation of outputs. Hereafter, we introduce aggregation uncertainty. In a price-based setup, aggregation uncertainty means that the prices are uncertain. Aggregation of input and outputs have been examined in several studies, for example Zelenyuk (2020), Simar & Zelenyuk (2018), Färe & Grosskopf (1985), Färe, Grosskopf, & Zelenyuk (2004). However, to the best of our knowledge, no one have studied the implications of uncertainty in these aggregations. Finally, we examine how this uncertainty influences the efficiency scores in four different cases.

In the first case, we calculate the changes in the efficiency scores in a general model with one input and two underlying cost-drivers, which are aggregated into one output. In the second case, we use the Danish waste water regulation to illustrate how aggregation uncertainty changes the efficiency scores empirically. In the third and fourth cases, we expand the empirical example to consider several variables with random noise in the underlying prices. We discuss whether the consumers or the companies should pay for the risk and modify the model based on this.

Hereafter we go into detail with a regression-based output aggregation scheme used by many European regulators, including DWRA, and show how these kinds of aggregations transform DEA from a non-parametric model to a more semi-parametric model, with close similarities to models such as Corrected Ordinary Least Square (COLS) and Stochastic Frontier Analysis (SFA).

The rest of this paper is structured as follows: Section 2 describes the DEA methodology. Section 3 introduces the implications of aggregating outputs and discuss the aggregation uncertainty. In section 4, we calculate and discuss the results from the Danish waste water regulation when we assume uncertainty in the aggregation scheme. Section 5 discusses a specific aggregation scheme based on regression analysis, and section 6 concludes the paper.

2 DEA methodology

DEA efficiency scores are estimated using linear optimization programs which can be interpreted in either the envelopment formulation or its dual multiplier formulation. The input orientated envelopment formulation with constant return to scale is given in (1)-(4).

$$
\begin{aligned}
&\min && \theta && && && (1)\\
&s.t. && \sum_{i=1}^{I} \lambda_i Y_i && \geq && Y_0 && (2)\\
& && \sum_{i=1}^{I} \lambda_i X_i && \leq && \theta X_0 && (3)\\
& && \lambda_i && \geq && 0 \quad \forall\, i && (4)
\end{aligned}
$$

The program minimizes the efficiency score, θ for the company under evaluation. Y_i is the output vector for company i, and X_i is the input vector. The index $i = 0$ indicates the company under evaluation. λ_i is a non-negative free variable used to calculate a convex combination of the peer companies.

The program in (1)-(4) has a dual formulation, called the multiplier formulation, shown in (5)-(8).

$$
\begin{aligned}
&\max && uY_0 && && && (5)\\
&s.t. && vX_0 && = && 1 && (6)\\
& && -vX_i + uY_i && \leq && 0 \quad \forall\, i && (7)\\
& && v, u && \geq && 0 && (8)
\end{aligned}
$$

In this program, we maximize the total value of the output for the company under evaluation, where u is a vector of output shadow prices (output multipliers) and v is a vector of input shadow prices (input multipliers). The efficiency score is the solution to the objective function, uY_0.

The two formulations yield the same efficiency score between 0 and 1, where 1 indicates a fully efficient company.

3 Implications of aggregation uncertainty in a general case

We can interpret output aggregations in DEA in two ways: In the envelopment formulation, the aggregation can be thought of as a fixed trade-off between the aggregated outputs. In the multiplier formulation, the aggregation can be interpreted as the total value of the aggregated outputs, where each output is valued equally across all the companies.

We illustrate an aggregation scheme in the envelopment formulation in Figure 1. We have four companies of which A, B, C are efficient, and D is inefficient, as indicated by the DEA frontier (black lines). The DEA frontier, in this example, consists of four facets. Each facet corresponds to a specific trade-off between the two outputs. In the multiplier formulation, each facet corresponds to a different set of output multipliers. This is, in fact, an aggregation being done inside DEA.

Now assume that the trade-offs/output multipliers (hereafter trade-offs) are incorrect. This can, for example, arise from the curse of dimensionality, because DEA gives the companies the benefit of the doubt, the implications of which gets more extreme as the dimensionality increases. In order to solve the problem, we can aggregate the two outputs prior to the DEA analysis, for example, based on the price approach. The green, red, and blue facets show three different aggregations, each using different prices. For a given set of prices, the corresponding facet now replaces the DEA technology set. We thereby go from four different trade-offs to assuming that all companies, in any given location within the technology set, have the same trade-offs between these two outputs.

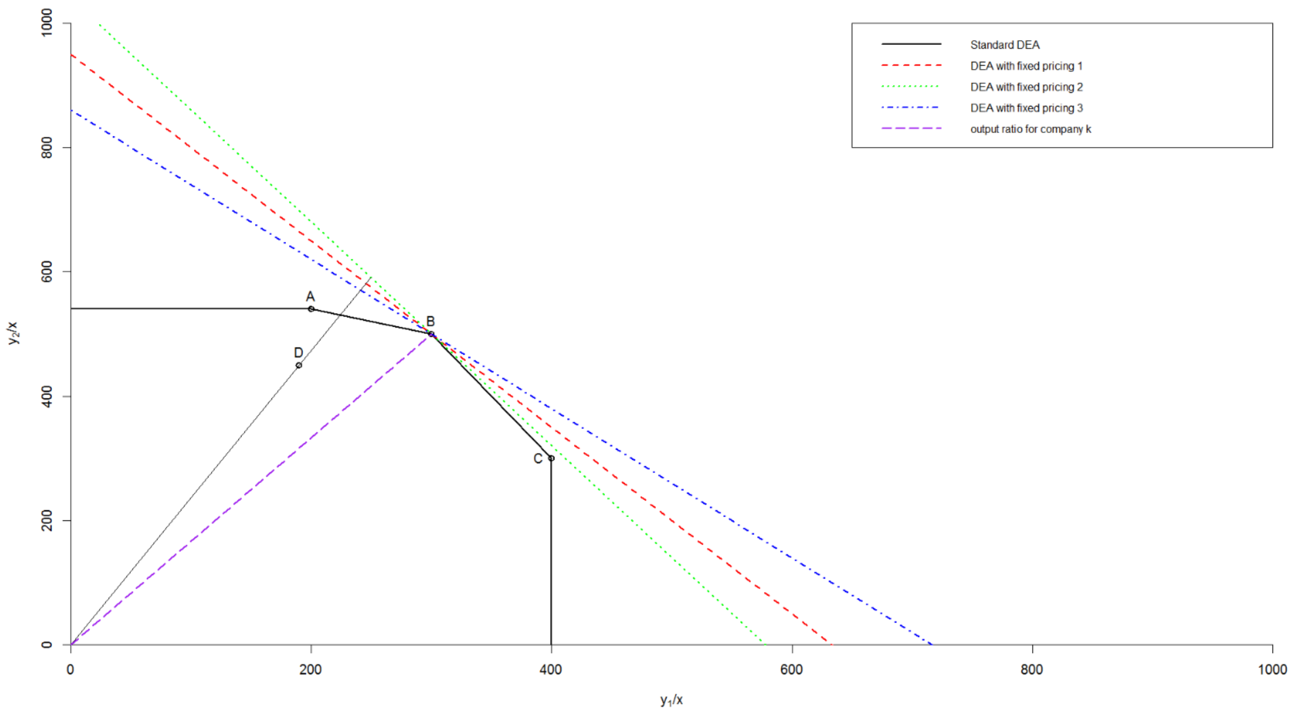

Figure 1 – DEA with four different aggregation schemes

Figure 1 shows how different aggregation schemes, with different sets of prices, influence the companies. We denote this as aggregation uncertainty. We observe that companies left of the purple line prefer the aggregation given by the blue facet and that companies on the right-hand side of the purple line prefer the aggregation of the green facet (disregarding the standard DEA frontier, which will always be preferable).

We will now show how this aggregation uncertainty influences the companies. For simplicity, we use only one input and two underlying cost-drivers, which we aggregate into one single output.

The input orientated multiplier DEA program with constant return to scale, one input (x), two outputs (y_1, y_2) and corresponding known output prices p_1, p_2, which we use to aggregate the outputs, is given in (9)-(12). v is the input multiplier, and u is the multiplier on the aggregated output.

$$\max \quad u(p_1 y_1^0 + p_2 y_2^0) \tag{9}$$

$$s.t. \quad vx^0 = 1 \tag{10}$$

$$-vx^i + u\left(p_1 y_1^i + p_2 y_2^i\right) \leq 0 \quad \forall\, i \in I \tag{11}$$

$$v, u \geq 0 \tag{12}$$

If we know company 0's peer[6], we can rewrite the DEA program as shown in (13), where f^0 is the efficiency score and superscript i^* is the peer. Note that we can multiply by x^0 to calculate the efficient costs instead of the efficiency score.

$$f^0 = \frac{x^{i^*}}{x^0} \cdot \frac{p_1 y_1^0 + p_2 y_2^0}{p_1 y_1^{i^*} + p_2 y_2^{i^*}} \tag{13}$$

The fraction $\frac{p_1 y_1^0 + p_2 y_2^0}{p_1 y_1^{i^*} + p_2 y_2^{i^*}}$ expresses the aggregated outputs for company 0 against company i^*. In other words, how much aggregated output one company produces compared to the other. It then also reflects the proportion of input needed for company 0. So, for example, $\frac{p_1 y_1^0 + p_2 y_2^0}{p_1 y_1^{i^*} + p_2 y_2^{i^*}} = 0.5$ means that company 0 produces half the output of company i^* and that it, therefore, should only use half the amount of input as well.

Now assume that we have an alternative set of prices, $\tilde{p}$ and corresponding efficiency scores, $\tilde{f}$. We can describe the aggregation uncertainty as the ratio between the efficiency scores from the two models:

$$\frac{f^0}{\tilde{f}^0} = \frac{\left(\frac{x^{i^*}}{x^0} \cdot \frac{p_1 y_1^0 + p_2 y_2^0}{p_1 y_1^{i^*} + p_2 y_2^{i^*}}\right)}{\left(\frac{x^{i^*}}{x^0} \cdot \frac{\tilde{p}_1 y_1^0 + \tilde{p}_2 y_2^0}{\tilde{p}_1 y_1^{i^*} + \tilde{p}_2 y_2^{i^*}}\right)} = \frac{\left(1 + \frac{y_1^0}{y_2^0} \cdot \frac{p_1}{p_2}\right)\left(1 + \frac{y_1^{i^*}}{y_2^{i^*}} \cdot \frac{\tilde{p}_1}{\tilde{p}_2}\right)}{\left(1 + \frac{y_1^{i^*}}{y_2^{i^*}} \cdot \frac{p_1}{p_2}\right)\left(1 + \frac{y_1^0}{y_2^0} \cdot \frac{\tilde{p}_1}{\tilde{p}_2}\right)} \tag{14}$$

This result gives us two insights:

1) The ratio between the efficiency scores in the two models is exclusively influenced by the company's output mix compared to that of its peer and the direction of the price ratio change. Therefore, it is not influenced by the company's input, level of output or efficiency.
2) If $\frac{y_1^0}{y_2^0} < \frac{y_1^{i^*}}{y_2^{i^*}}$, company 0 will prefer that $\frac{p_1}{p_2} < \frac{\tilde{p}_1}{\tilde{p}_2}$ as this gives them a higher efficiency score, as evident from $\frac{f^0}{\tilde{f}^0} > 1$. If $\frac{y_1^0}{y_2^0} > \frac{y_1^{i^*}}{y_2^{i^*}}$ the company will instead prefer $\frac{p_1}{p_2} > \frac{\tilde{p}_1}{\tilde{p}_2}$ as that will too result in $\frac{f^0}{\tilde{f}^0} > 1$. This corresponds to the intuition discussed in Figure 1.

It is not trivial to deduce from equation (14) how much the efficiency scores change given a change in the prices, as it depends on how different the companies are in terms of their output mixes. If the companies operate with very similar output mixes, the prices do not matter and, therefore, the aggregation uncertainty can be ignored. However, if the companies operate with considerable differences in the output mixes, the aggregation uncertainty gets very important.

To better understand these results, we calculate and discuss the size of $\frac{f^0}{\tilde{f}^0}$ for the Danish waste water regulation in the next section.

[6] Note that there only is one peer in a DEA model with one input, one output and constant return to scale.

4 Example – The Danish waste water regulation

To illustrate the implications of aggregating two outputs, we use the benchmarking model from the Danish waste water regulation as an example. DWRA uses costs as input and OPEX and CAPEX net volumes as outputs in this model. The model is input orientated, and DWRA assumes CRS. We illustrate the model in Figure 2. where the solid black line indicates the standard DEA frontier.

If we examine the facet structure in this model, applying the convex hull algorithm of Petersen & Olesen (2015), we identify the three efficient facets given by the normal vectors in Table 4.1. These correspond to marginal rates of substitution (MRS) between OPEX and CAPEX of 0, 2.75 and Infinity, respectively, for the three facets. Because both OPEX and CAPEX are measured in DKK, one could argue that the MRSs should be 1.[7]

Table 4.1 – Normal vectors and marginal rates of substitution. Note that the offsets to the normal vectors are omitted because they are all zero in crs

	Costs	OPEX	CAPEX	$\partial OPEX/\partial CAPEX$
Facet 1	-0.7242737	0.0000000	0.6895125	0
Facet 2	-0.5719557	0.7710187	0.2799944	2.753694
Facet 3	-0.5085658	0.8610231	0.0000000	Infinity

If we force the model to have MRSs between OPEX and CAPEX equal to one, this corresponds to aggregating the two outputs with equal prices. We illustrate this with the green facet in Figure 2. Ignoring the standard DEA frontier and using this green facet as the new frontier instead results in lower efficiency scores for all companies, except for the companies positioned precisely on the purple line, which indicates the now only efficient company's output mix[8]. As we discussed in section 3, the further away from the purple line a company is, the bigger the changes in efficiency scores.

[7] This requires that the companies have an actual trade-off between OPEX and CAPEX, but for now, we do not question this assumption.

[8] In this example, no other companies are positioned precisely here.

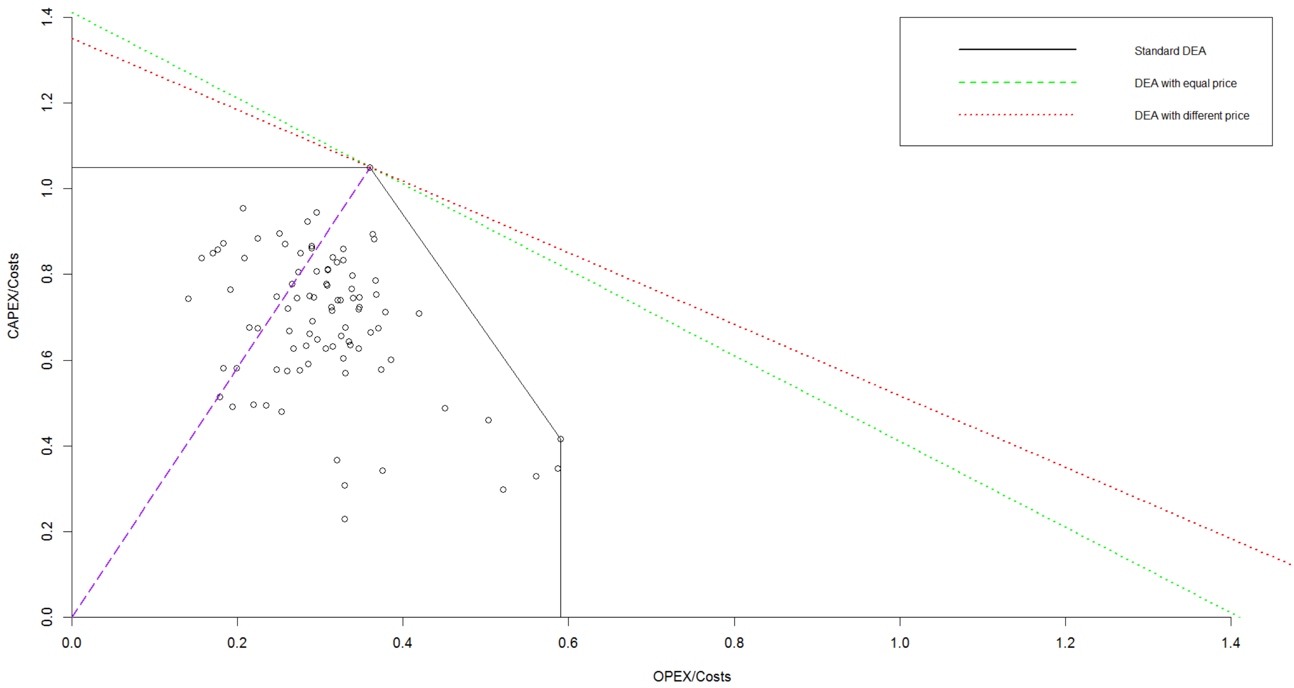

Figure 2 – A standard DEA frontier and two alternative frontiers based on different aggregations of outputs

Now assume, as is arguably the case in the Danish regulation, that this aggregation scheme is questionable because the two outputs are calculated using different techniques, by different consultants, and at different times. One OPEX DKK is therefore not necessarily the same as one CAPEX DKK. To examine the implications of this, we first calculate the efficiency scores using equal prices and hereafter the efficiency scores in a model where the price on CAPEX is 1.2 times higher than the price of OPEX. In other words, we examine an uncertainty on the CAPEX price at 20 % as an upper bound and 0 % as the lower bound. In Figure 2, this corresponds to using the red frontier as an upper bound and the green frontier as a lower bound.

Figure 2 shows a clear difference in the technology when we go from the standard DEA model with two outputs to the one with equal prices between the two outputs. The changes will be relatively small for companies close to the purple line (indicating the peer unit's output mix). For companies further away, the difference gets quite big. The difference between the technology with equal prices and the technology, where the price of CAPEX is 1.2 higher than the price of OPEX, is much smaller. We show how the efficiency scores change in Figure 3.

Figure 3 shows the companies' output mix, $\frac{OPEX}{CAPEX}$ on the horizontal axis and the efficiency score ratio between the two models, $\frac{f}{\tilde{f}}$ on the vertical axis. We define f as the efficiency scores calculated with the green facet in Figure 2 and $\tilde{f}$ as the efficiency scores calculated using the red facet. If $\frac{f}{\tilde{f}} = 1$, the company gets identical efficiency scores in the two models, which in this case only happens for the one efficient company. We observe that most companies have an efficiency score ratio close to 1, meaning that the aggregation uncertainty is not severe for these companies. A few companies with a high output mix have a ratio around 0.95 and are therefore more dependent on the prices chosen to aggregate the outputs.

We will argue that the companies in the Danish waste water sector have so similar output mixes that an aggregation uncertainty of 20 % does not have extreme consequences. However, in a regulatory context, even a slight decrease in the efficiency score will still be costly for the companies, and it is, therefore, essential for the regulator to minimize this uncertainty.

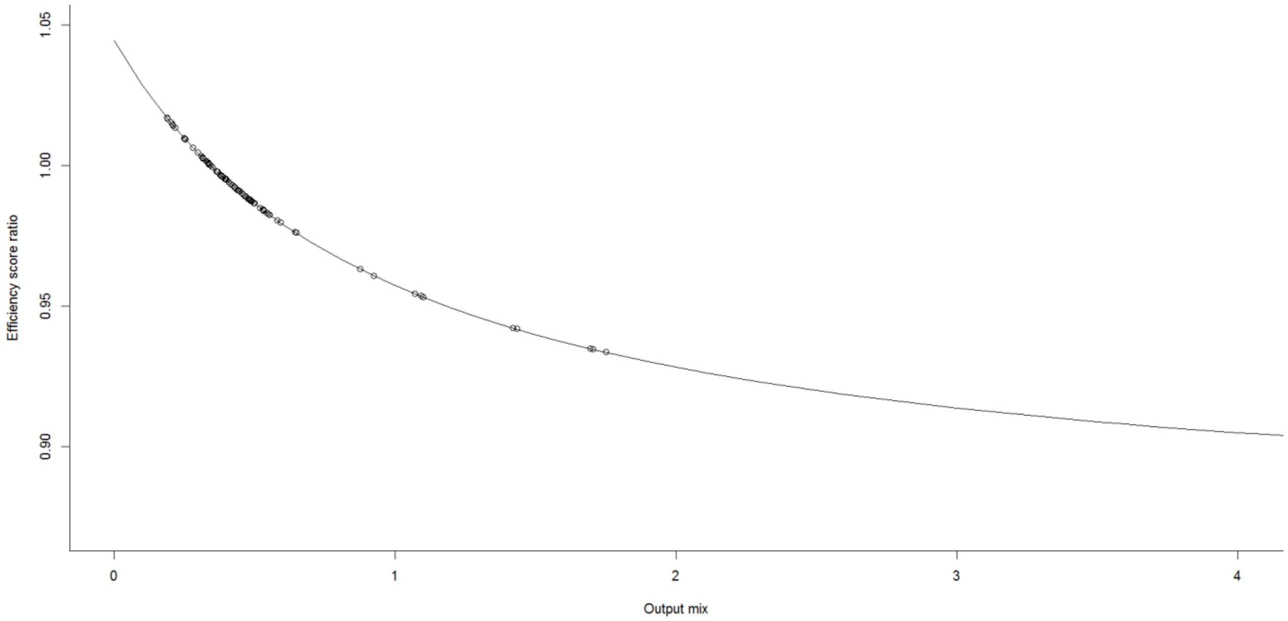

Figure 3 – Relationship between the companies output mix and their efficiency score ratio between the two models

In the most extreme cases, if a company has only one of the two outputs, the efficiency score ratio will be 1.045 for an output mix of $\frac{OPEX}{CAPEX} = 0$ and 0.870 for $\frac{CAPEX}{OPEX} = 0$. If such companies existed, the precise aggregation scheme would be more critical depending entirely on the efficient company's output mix.

As mentioned in section 3, the efficient company's output mix is essential for the efficiency score ratio between the two models. Therefore, we simulate new output mixes for the efficient company and in Figure 4 show each iteration's efficiency score ratio function (equation (14)). We let the output mix for the efficient company $\left(\frac{OPEX^{i^*}}{CAPEX^{i^*}}\right)$ vary between 0.1 and 1 with increments of 0.1. Note that the efficient company's output mix lies where the efficiency score ratio equals 1.

We observe that the function is offset towards the northeast when the output mix for the efficient company increases. Otherwise, the tendency is the same; companies with similar output mixes are not exposed to aggregation uncertainty, while companies with an output mix fare from that of the efficient company experience big changes in the efficiency score due to new aggregation prices.

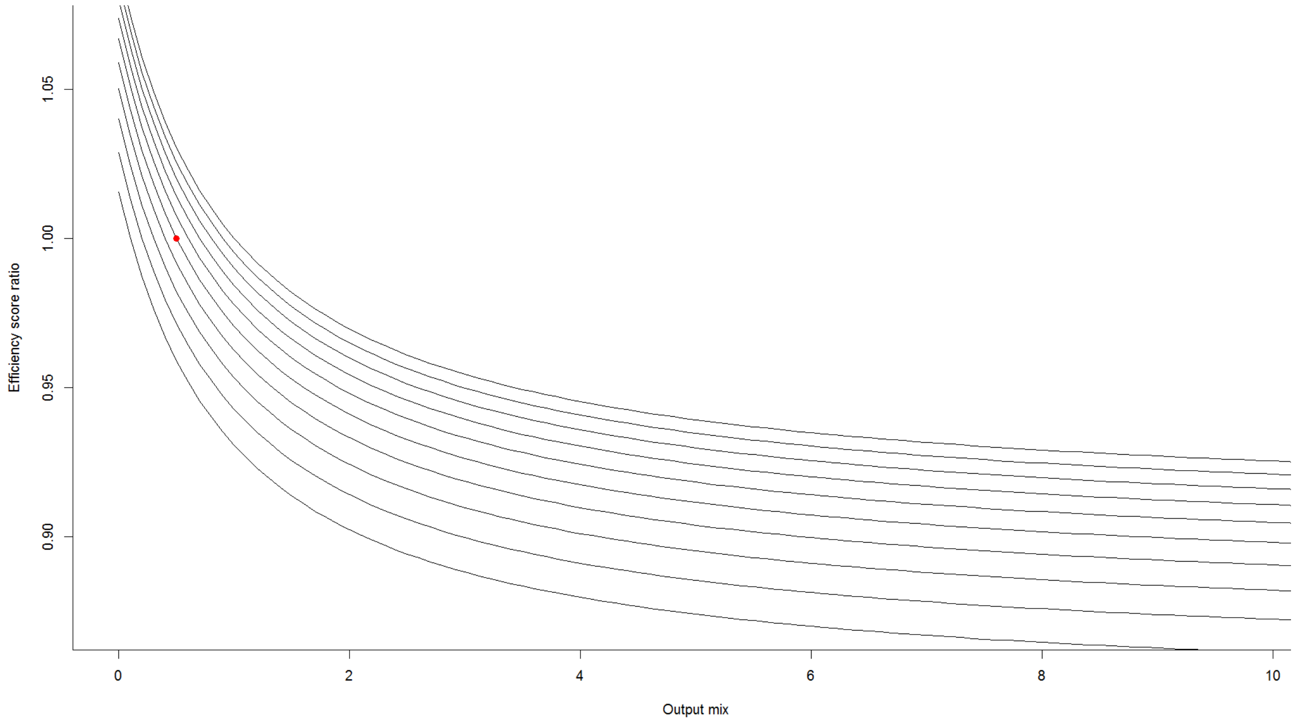

Figure 4 - Relationship between the companies' output mixes and their efficiency score ratio between the two models for 10 different output mixes for the efficient company. For example, the fourth lowest line shows the relationship when the efficient company's output mix is 0.5, indicated by the red dot

We use a permutation test to examine whether the aggregation uncertainty significantly influences the results (Asmild, Kronborg, & Rønn-Nielsen, 2018). We describe the test in appendix A. The test is divided into two parts. First, we test whether the frontier gets significantly better (the technology set is expanded) when we increase CAPEX prices by 20 %, corresponding to the red frontier in Figure 2. We find that the frontier gets significantly better. However, this is no surprise because an increase in one output price, while holding everything else equal, results in an upward parallel shift of the frontier.

Second, the permutation test examines if the distribution of the efficiency scores changes. We do not find evidence for a significant change in the distribution. This result occurs because the losses/gains in the efficiency scores approximately cancel out between the companies. We do not test if there is a significant change in the efficiency score for the individual companies.

We have shown how the aggregation of two outputs influences the efficiency scores in the Danish waste water regulation. The following sections examine three different aggregation scenarios – all for the same data. First, in section 4.1, we split the CAPEX into two parts to examine the aggregation between two outputs when we have three outputs. Then, in section 4.2, we disaggregate CAPEX into all its underlying cost-drivers and simulate the CAPEX aggregations introducing random noise on the prices. Lastly, in section 4.3, we use weight restrictions on the underlying cost-drivers for CAPEX to give the companies the benefit of the doubt regarding the aggregation uncertainty.

4.1 Aggregation of a subset of the outputs

Standard DEA models often have more than two outputs, making it more complicated to generally interpret $\frac{f^0}{\tilde{f}^0}$ which no longer only depends on the output mixes but also the dual multipliers. In this section and sections 4.2 and 4.3 below, we split the CAPEX net volume into multiple outputs to examine the changes in the efficiency scores in models with several outputs. In this section, we split CAPEX into the distribution and production processes, where the production process is cleaning and disposing of the waste water. Doing so, we can examine the consequences if the prices in the distribution process are over/underestimated compared to those of the production process.

In addition, we no longer assume that OPEX and CAPEX can be aggregated. We, therefore, have three outputs (OPEX, CAPEX distribution and CAPEX production) where we want to aggregate the two CAPEX measures introducing uncertainty of 20 % following the logic in section 3.

Note that the changes in the efficiency scores still mainly depend on the companies' output ratios, but now we also need to consider the dual multipliers. If, for example, the multiplier for OPEX is 0, the changes in the efficiency scores still only depend on the output mix for the two aggregated outputs. However, as the multiplier for OPEX increases, the changes in the efficiency scores decrease. In other words, the aggregation uncertainty between two outputs is only relevant if the two outputs are actually used in the benchmark (have positive multipliers), and the higher the multipliers on these outputs, the more important is the uncertainty.

The results are shown in Figure 5. We observe that the changes in the efficiency scores are smaller than in the model with only two outputs. This is expected because the aggregation uncertainty only is relevant for a subset of the variables, and we thereby implicitly assume that there is no aggregation uncertainty on the last variable(s). In addition, we observe a single company with a low ratio of 0.954, meaning that they are sensitive to the aggregation uncertainty. This is because this company has almost zero CAPEX distribution[9], which differs a lot from its peers.

[9] It has almost zero CAPEX distribution because it is collaborating with its neighbouring waste water companies such that the other companies are in charge of the distribution, and this company is in charge of the production.

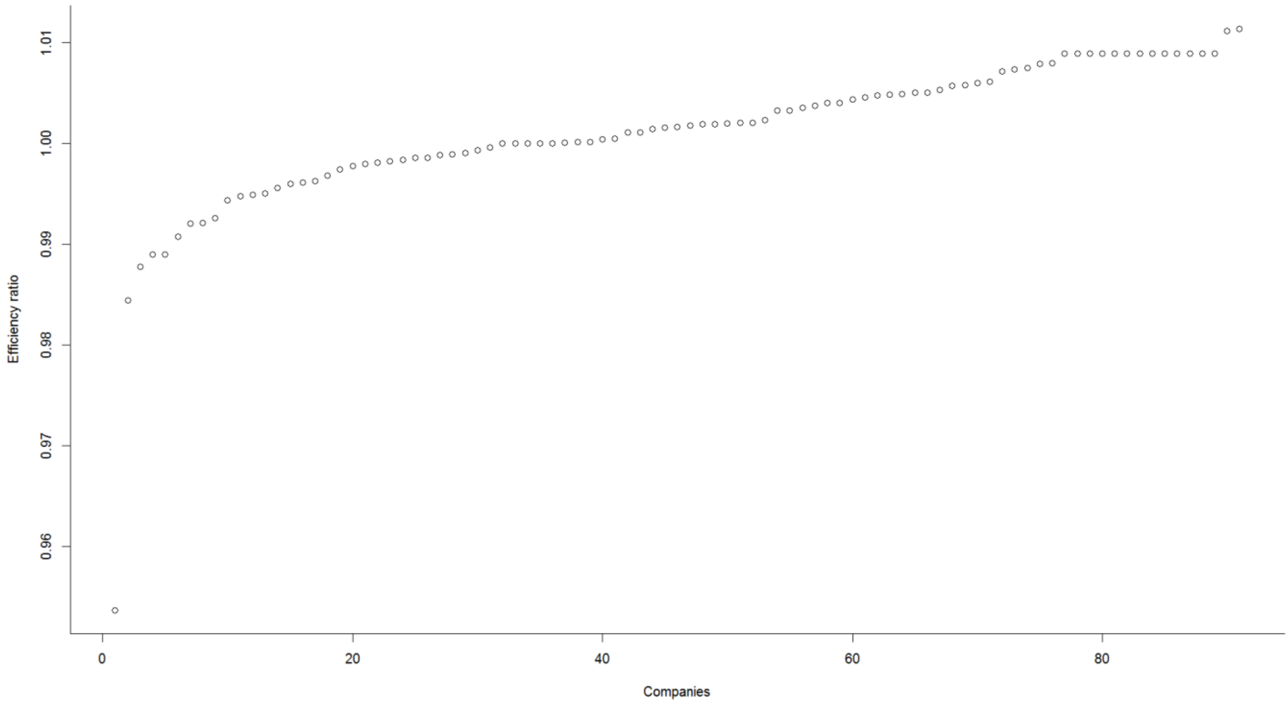

Figure 5 – Efficiency score ratio between the models with different underlying aggregation prices

Following the logic from section 4 above, we test whether the aggregation uncertainty significantly influences the results, using the method described in appendix A. First, we find that the frontier gets significantly better. The argumentation is the same as in section 4: an increase in one price while holding everything else equal results in an upward parallel shift of the frontier.

Second, we do not find evidence for a significant change in the distribution of the efficiency scores. Again, the argumentation is the same as in section 4; the losses/gains in the efficiency scores approximately cancels out among the companies.

To conclude this section, we find that the Danish waste water companies are relatively homogenous concerning their output mix, so the aggregation uncertainty does not matter a lot in this setup.

4.2 Risk as random variation among underlying cost drivers

In this section, we examine the aggregation uncertainty for each of the underlying prices of CAPEX[10] by creating new CAPEX net volumes with random noise around the current prices. We thereby have two outputs; OPEX and a modified CAPEX.

We use the following iterative process:

1) We change the prices, P, separately using a random risk factor, r, from a normal distribution with a standard deviation of 0.1: $\hat{P}_i = P_i \cdot r[0,0.1], \ \forall\, i \in [CAPEX_I]$ where $CAPEX_I$ is a vector of all the underlying CAPEX cost-drivers

[10] A portion of the underlying CAPEX outputs has a non-linear price structure, which we for simplicity do not change here

2) We create the new CAPEX net volume based on the estimated prices: $\hat{y} = CAPEX_I \cdot \hat{P}_I$
3) We run DEA where the new net volume $\hat{y}$ replace CAPEX and report back the results

We repeat step 1:3 10,000 times. The results are shown in Figure 6. The figure shows the companies' efficiency scores for each iteration. We observe that there are two efficient companies (in the upper right corner), which are efficient in all iterations. Following equation (14), the range and density of the efficiency scores for each company follow their output mixes compared to those of the two efficient companies[11]. However, some companies have a high multiplier weight on OPEX, which means that their efficiency range is small even with a different output mix. On the right-hand side of Figure 6, we observe three companies with no range (in addition to the two efficient companies), meaning that they have no multiplier weights on CAPEX and that the aggregation uncertainty within the simulated interval is not relevant[12].

The companies have an average efficiency range of 0.068. In the most extreme case, a specific company's efficiency range is 0.166.

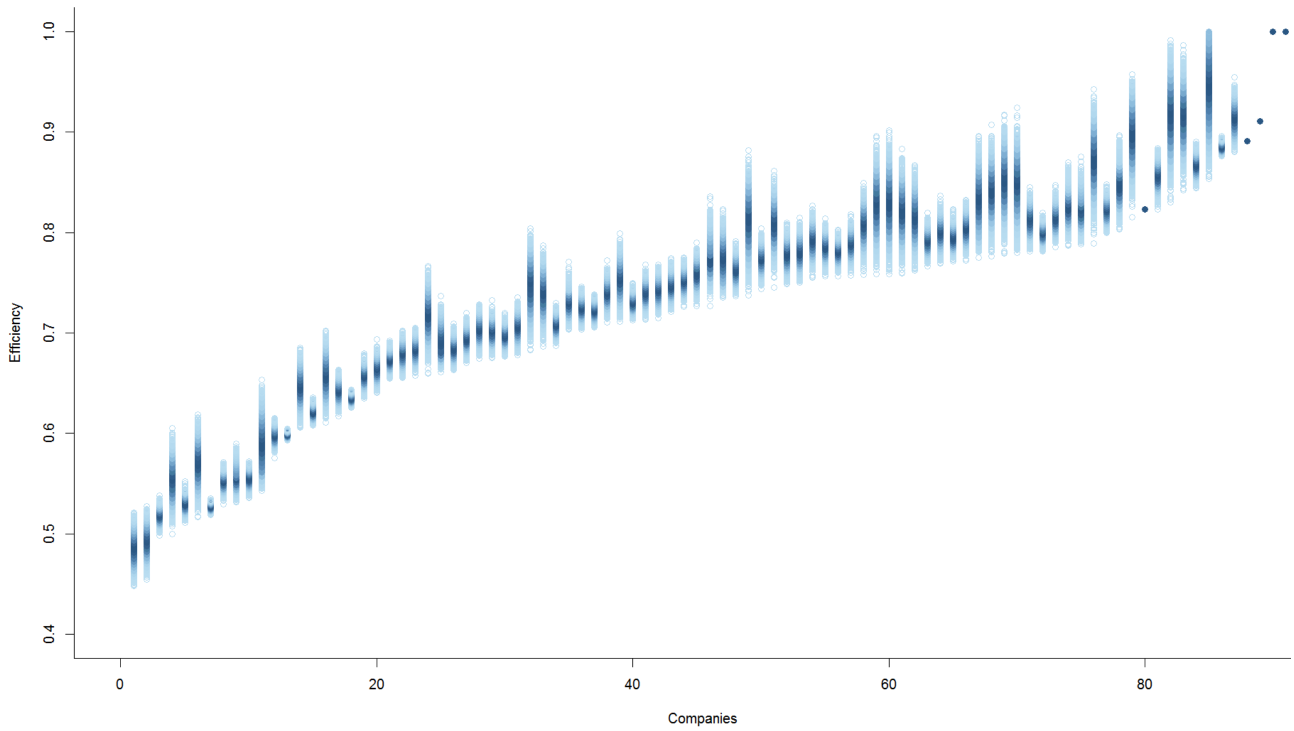

Figure 6 – Efficiency density for all companies when introducing aggregation uncertainty. The dark colours indicate a high density

In Figure 7, we zoom in on a random company to better examine the efficiency density. Most companies have the same density structure. The red cross shows the initial efficiency score. We observe that the initial efficiency score is in the middle of the range and that the density seems normally distributed around this. In most scenarios, the company's efficiency score will not change substantially, but we observe drastic changes in a few unlikely scenarios.

[11] Some companies only have one peer. In that case, only that peer is relevant.
[12] If we increase the standard deviation, we could potentially observe that these companies, in some situations, would have a positive multiplier weight on CAPEX.

This indicates that most combinations of prices around the original price do not notably change the efficiency score – an increase in some prices favours this company while a decrease in other prices does not and vice versa. However, in the extreme iterations, we find a random set of prices that is exclusively good for this individual company. In other words, we randomly change the prices for the underlying cost-drivers for which this company is unique compared to its nearest peers.

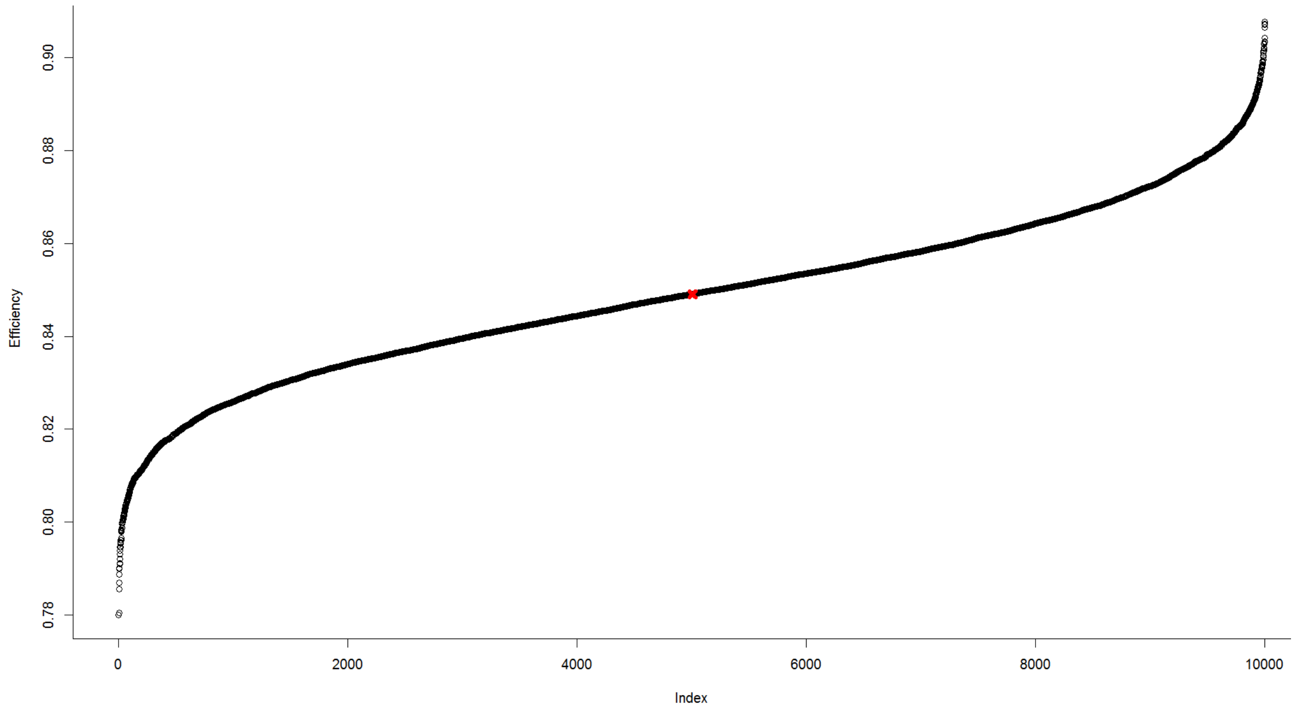

Figure 7 – Efficiency density for a random but representative company

On average, the efficiency scores across the iterations are very similar to the initial efficiency score for most companies. For the company with the biggest difference between the average efficiency score and the initial efficiency score, the average is only 0.2 % higher than the initial. The average efficiency score is 0.02 % lower than the initial model for the company with the largest reduction.

However, as we showed earlier, the prices in some iterations make considerate changes to the efficiency scores for individual companies if the chosen prices favour this specific company. If the chosen prices favour the efficient companies, it will lead to a general decrease in most inefficient companies' efficiency scores. If the chosen prices are unfavourable for the efficient companies, it will lead to a general increase instead. As a result, the models average efficiency score changes across the iterations with a maximum of $+2.7$ % and -2.5 % from the original average.

To test whether the changes in the frontier and efficiency scores are significant, we continue using the statistical test described in appendix A. However, this requires a permutation test for each of the 10,000 iterations used in this section, resulting in 10,000 p-values[13]. The p-values indicate whether the frontier shift and efficiency change between the initial model and the model for a given iteration is significant. We report the p-values in Figure 8.

[13] Due to limited computing power, we reduce the number of iterations to 1,000

The figure shows considerable differences between the iterations. The frontier shift is significant in a few iterations but insignificant in most. This means that the uncertainty in the aggregation prices can result in a significant frontier shift, but in most cases, it will not.

The distribution of the efficiency change shows that most p-values are in one of the two tails. This is because the efficiency change depends a lot on the few efficient companies (we observe either one or two efficient companies in the iterations). Therefore, the results depend on how much the most relevant prices for the efficient companies change.

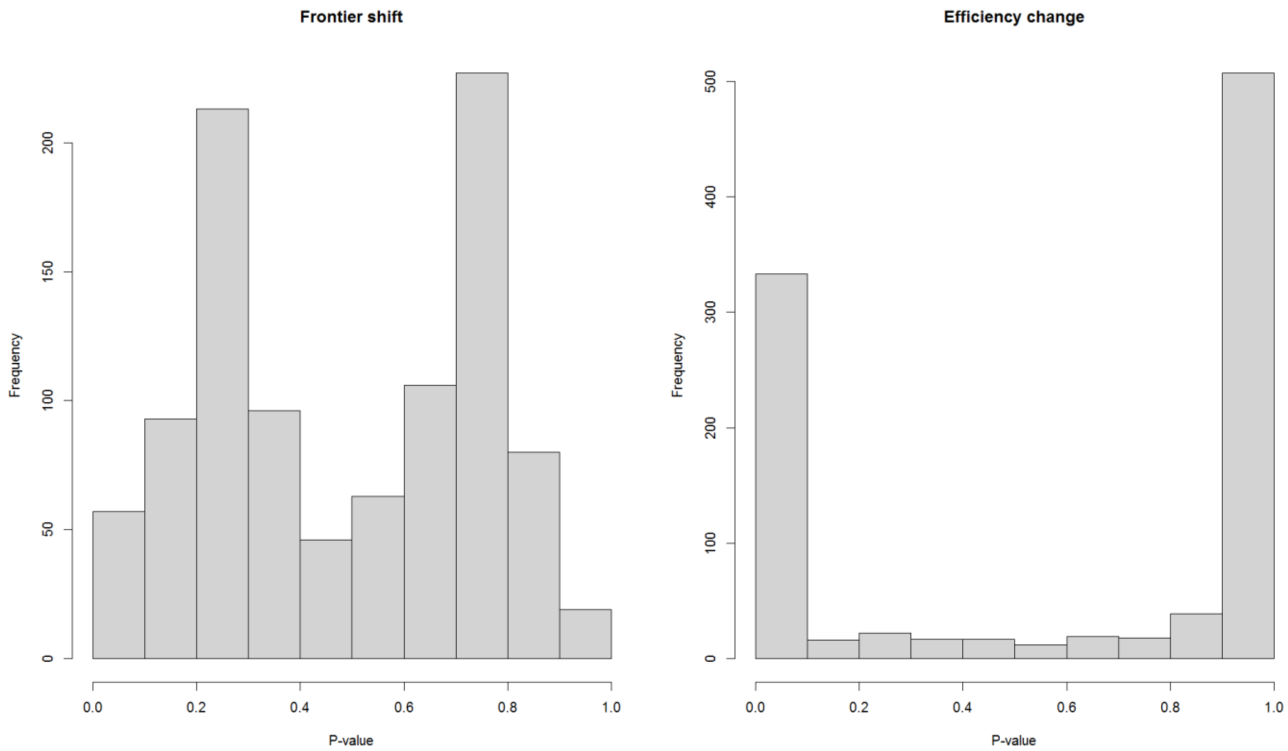

Figure 8 - P-values from the permutation tests

To conclude this section, we find that random changes in the prices do not substantially change the average companies' efficiency even while the changes are significant in some iterations. However, for individual companies, the price risk can be severe. Knowing this, the regulator needs to decide if it is fair (from the companies' point of view) that the companies are subject to such high risks solely based on model technicalities, which they cannot control themselves. Therefore, we transfer this aggregation risk from the companies to the consumers in section 4.3 below.

4.3 Risk in a benefit of the doubt setup

Section 4.2 above examined the companies' aggregation uncertainty concerning random noise in the underlying aggregation prices. The regulator's goal is to minimize consumer prices by setting efficiency requirements for the companies. However, random noise must not lead to these efficiency requirements being too high, as it in extreme consequence can lead to bankruptcy, which is not in anyone’s best interest. In the model from section 4.2, the consumers and companies split the risk, as we assumed a normally distributed noise term with a mean of zero. One can instead argue that the aggregation uncertainty should be paid solely by the consumers to minimize the risk of bankruptcy. This section uses weight restrictions based on the aggregation prices to do precisely this.

We use the same basic model as the previous section with costs as input and OPEX and CAPEX as outputs. However, instead of simulating new CAPEX net volumes with a random noise term on the prices, we now include all the underlying CAPEX cost-drivers as individual output constraints with weight restrictions based on the prices.[14]

For each underlying CAPEX cost-driver, we can add the weight restriction given in (15) where the $p's$ are the prices, u is the output multipliers, and i count the underlying cost-drivers in $CAPEX_I$

$$\frac{p_i}{p_1}u_1 - u_i = 0, \qquad \forall\, i \in [CAPEX_I] \tag{15}$$

Doing so yield the same results as using the net volumes. However, now we can manipulate the weight restrictions to contain risk. One way of doing so is changing the weight restriction to the following:

$$\left(\frac{p_i}{p_1} \cdot (1 \pm r)\right)u_1 - u_i = 0, \qquad \forall\, i \in [CAPEX_I] \tag{16}$$

where $r = 0.2$ is the maximum possible uncertainty in the ratio between the prices. We use 264 cost drivers, which gives 528 weight restrictions.

Using these weight restrictions, we allow the companies to choose the prices (in a given interval around the default prices) that gives them the best outcome. In other words, we remove all the aggregation uncertainty from the companies. In the simple setup from Figure 2, this corresponds to using the red facet on the left side of the efficient unit and then switching to the green facet on the right side of the efficient unit. For a more detailed link between the multiplier and envelopment formulation see Podinovski (2004). We show the results in Figure 9.

Figure 9 compares the companies' efficiency scores in the two models: DWRA's original model with fixed net volumes and the new model with 20 % aggregation uncertainty measured with weight restrictions. The diagonal line indicates equal efficiency scores in the two models, which is only the case for a few inefficient companies besides, of course, the original efficient companies. We observe that all other companies are above this line, meaning that they get a higher efficiency score when we introduce aggregation uncertainty as weight restrictions. In addition, we observe quite large changes in the efficiency scores, with the most drastic increase being 0.26. The gains from the weight restrictions seem a bit lower for the most inefficient companies, but overall we do not observe any clear patterns.

[14] A portion of the underlying CAPEX cost drivers has a non-linear price structure, which we, for simplicity, do not change with weight restrictions. Instead, these outputs are added together with their original prices, and the sum is hereafter included using the same principles as the remaining underlying cost-drivers.

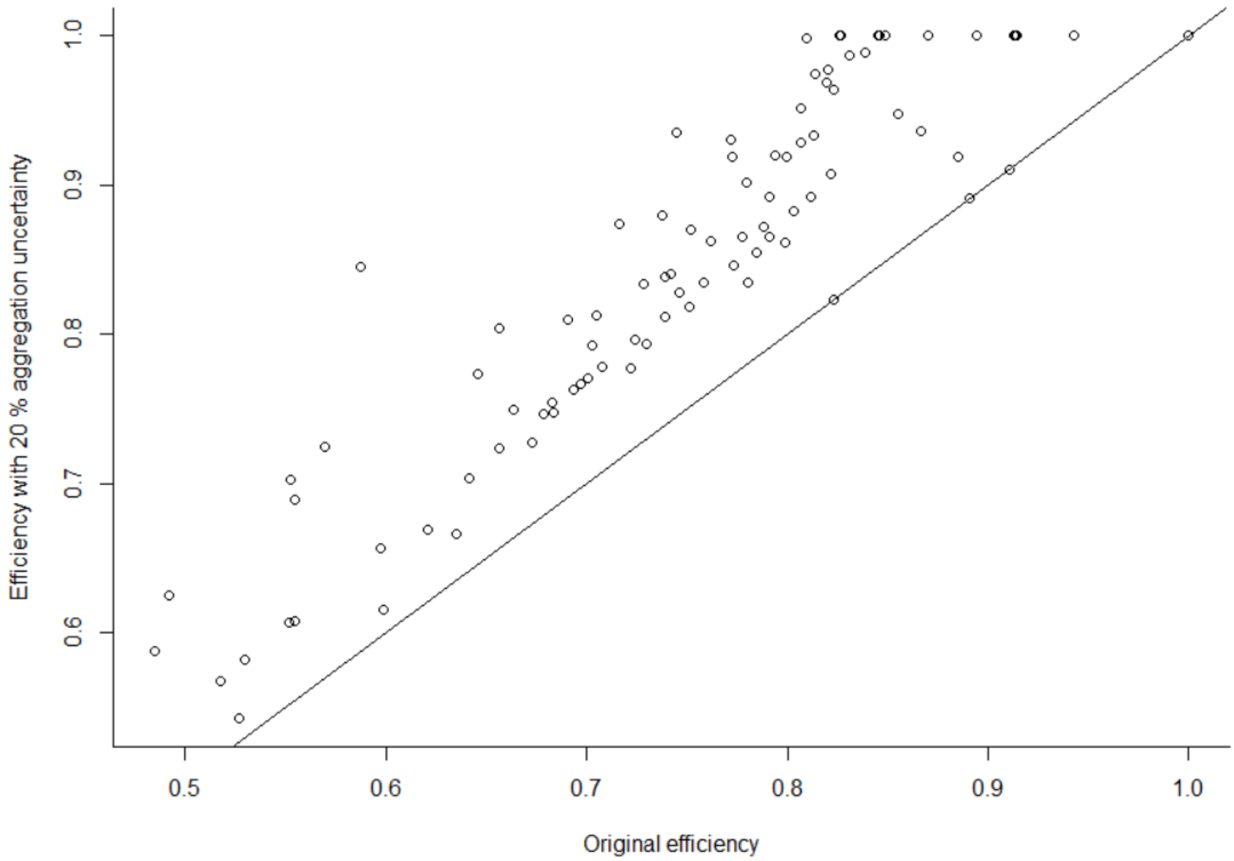

Figure 9 – Comparison of original efficiency and efficiency with 20 % aggregation uncertainty using weight restrictions

We show the summary statistics for the efficiency scores in the two models in Table 4.2. The table shows that the companies, on average, get a 10 % increase in their efficiency scores in the new model. This gain is equally distributed across the quartiles, which is consistent with our discussion in equation (14), where we showed that the current level of efficiency did not influence the efficiency changes when the aggregation prices change. The number of efficient companies goes from 2 to 13.

Table 4.2 – Summary statistics for the efficiency scores in the two models

Model	**Min.**	**1st Qu.**	**Median**	**Mean**	**3rd Qu.**	**Number of efficient companies**
Original	0.4848	0.6833	0.7622	0.7477	0.8209	2
With weight restrictions	0.5422	0.7690	0.8613	0.8436	0.9358	13

Lastly, Table 4.3 shows that most weight restrictions are binding. The table shows that 218 underlying CAPEX cost-drivers have zero companies with non-binding corresponding weight restrictions. In other words, all the companies set the prices at either the maximum or minimum allowed price for these outputs. Twenty-six outputs have a single company with non-binding weight restrictions, and finally, we observe that one specific output have 29 companies with a corresponding non-binding weight restriction.

Table 4.3 thus shows that the companies want to change most of the prices as much as possible in a specific direction inside the 20 % interval around the initial prices. This means that the companies are vulnerable to the aggregation prices chosen by the regulator and that it will be problematic for the sector to agree on a set of true prices. The frontier calculated in this model is thereby mostly calculated based on these prices and not the standard DEA axiom of convexity between the companies. In addition, if the return to scales is assumed anything but constant, these binding weight restrictions will somewhat overrule the assumption about the returns to scale and impose a constant return to scale-like assumption instead.

Table 4.3 – Frequencies of non-binding weight restrictions

Number of underlying CAPEX cost-drivers	218	26	6	4	1	3	3	1	1	1
Number of companies with non-binding restrictions	0	1	2	3	4	5	6	13	17	29

As in the previous sections, we test whether there are significant differences between the two models. Because we no longer have two sets of inputs and outputs, we need a new statistical test. The new test is based on the work of Rønn-Nielsen, Kronborg, & Asmild (2019), which test if there is a significant difference between the CRS and VRS assumption in a model. We describe the test and how we modify it in Appendix B.

Note that we can reformulate DWRA's initial model by splitting up CAPEX and using weight restrictions corresponding to (15). By doing this, we assume that one monetary unit has the same value among all the cost-drivers. In the new model with aggregation uncertainty, we expand this assumption such that the value of one monetary unit can vary with $\pm$ 20 %, cf. equation (16). Therefore, the model with aggregation uncertainty is nested into the initial model.

In the test described in Appendix B, we test the hypothesis that one monetary unit have the same value among all the cost-drivers within the assumptions of the nested model.

The permutation test gives a p-value of 0.182. This indicates that we cannot reject the initial model used by the Danish waste water regulator.

To conclude this section, we will argue that the companies are vulnerable to the aggregation prices chosen by the regulator, even while we cannot statistical reject the model. We observed considerable changes in efficiency when we removed all the risks from the companies and observed that most of the weight restrictions were binding. These extreme results shall, of course, be seen in the light of a relative high aggregation uncertainty of 20 %, and it is doubtful that the prices chosen by the individual companies, given the model specifications, are, in fact, the true prices. The companies chose the prices that yield the highest possible efficiency score, and the results need, therefore, to be interpreted as a precautionary measure for the benefit of the companies.

There is a close relationship between the methods used in this section and section 4.2 above. In section 4.2, we let the prices vary randomly such that the companies, on average, did not change their efficiency score. This section took the most extreme set of prices possible, given an aggregation uncertainty of 20 %, for each company. These prices are, in other words, just an extreme draw from the method used in the previous section.[15]

5 Regression-based output aggregations

We have so far assumed that the aggregation prices are known - at least approximately within some interval. In this section, we examine how the Danish waste water regulator (DWRA) and many other European regulators actually calculate the prices.

DWRA uses two different methods for OPEX and CAPEX. However, both methods are similar in that they calculate input prices instead of output prices – even while these are used for output aggregation. The model interpretation is, thus, how much the companies actually spend compared to the expected costs. This interpretation makes sense because the net volumes are assumed to be fixed - the companies can, for example, not change the number of customers or the demand. At the same time, DWRA assumes that the companies do not buy capacity that they do not need. Therefore, the model only examines if the companies are cost-efficient and not if they buy the correct assets (output).

In this paper, we focus on the calculation of OPEX because the CAPEX prices are calculated by external consultants without much documentation. Therefore, we pretend that the consultants use the same method for CAPEX as the regulator does for OPEX.

The prices are calculated using regression analysis in (17). However, due to the high number of cost drivers, DWRA in practice split the regression into several regressions and add the results afterwards, something which we ignore here.

$$x_i = f_i(\boldsymbol{q_i}, \boldsymbol{\beta_i}) + \epsilon_i, \qquad \forall\, i \in [OPEX, CAPEX] \tag{17}$$

Here x is the costs, $\boldsymbol{\beta}$ are the coefficients, $\boldsymbol{q}$ are the underlying cost drivers, and ϵ is the error term. The index i indicates whether we calculate OPEX or CAPEX. The net volumes are then defined as the fitted values from (17):

$$y_i = f_i(\boldsymbol{q_i}, \boldsymbol{\beta_i}), \qquad \forall\, i \in [OPEX, CAPEX] \tag{18}$$

Here y_i is the OPEX and CAPEX net volume, respectively. The only difference between the input (x) and outputs (y) is, thus, given by the error term. If we insert this in the DEA model, the companies are compared according to how big an error term they have compared to the companies with the smallest error term relative to their size. We can show this by rewriting $y_i = x_i - \epsilon_i$ and inserting this in the input orientated DEA multiplier program with a constant return to scale as follows:

The standard DEA multiplier program is given in (11)-(12).

---

[15] This requires that we censor the normally distributed risk factors such that they do not exceed $\pm$ 20%

$$\max \frac{\sum_{i=1}^{2} v_i y_i^0}{u\left(\sum_{i=1}^{2} x_i^0\right)} \tag{19}$$

$$\frac{\sum_{i=1}^{2} v_i y_i^k}{u\left(\sum_{i=1}^{2} x_i^k\right)} \leq 1, \qquad \forall\, k \tag{20}$$

Here u is the multiplier for the output, v is the multiplier for the input and the index k count the companies. $k = 0$ is the company under evaluation. Note that the costs are added together to a single input cf. the discussion in section 0. We rewrite $y_i = x_i - \epsilon_i$ and insert in the DEA program:

$$\max \frac{\sum_{i=1}^{2} u_i (x_i^0 - \epsilon_i^0)}{v\left(\sum_{i=1}^{2} x_i^0\right)} \tag{21}$$

$$\frac{\sum_{i=1}^{2} u_i (x_i^k - \epsilon_i^k)}{v\left(\sum_{i=1}^{2} x_i^k\right)} \leq 1, \qquad \forall\, k \tag{22}$$

We know that the condition will be binding for at least one k. We can, therefore, calculate the efficiency score in (23), where k indicates a company that is a peer for the company under evaluation.

$$f_{DEA} = \frac{\dfrac{\sum_{i=1}^{2} u_i (x_i^0 - \epsilon_i^0)}{v\left(\sum_{i=1}^{2} x_i^0\right)}}{\dfrac{\sum_{i=1}^{2} u_i (x_i^{k^*} - \epsilon_i^{k^*})}{v\left(\sum_{i=1}^{2} x_i^{k^*}\right)}} = \frac{\sum_{i=1}^{2} x_i^{k^*}}{\sum_{i=1}^{2} x_i^0} \cdot \frac{\sum_{i=1}^{2} u_i \left(x_i^0 - \epsilon_i^0\right)}{\sum_{i=1}^{2} u_i \left(x_i^{k^*} - \epsilon_i^{k^*}\right)}, \qquad for\ k^* = peer \tag{23}$$

Thus, the efficiency score is calculated as the relative error terms $\left(x_i^0 - \epsilon_i^0\right)$ compared to the peers' relative error terms $\left(x_i^{k^*} - \epsilon_i^{k^*}\right)$ weighted together with the multipliers u_i. Therefore, the regulators' so-called non-parametric DEA model is more parametric than first assumed. This is problematic because one of the main arguments for using DEA is precisely that it is non-parametric. The question is thus; why use DEA instead of, for example, COLS and SFA?

To investigate this, we illustrate examples of the cost function in COLS, SFA and OLS in Figure 10. Note that the coefficients are the same in OLS and COLS because COLS is just a downwards offset of OLS. The figure illustrates SFA with the same coefficients as OLS and COLS. Empirically this is not necessarily correct because SFA considers noise and inefficiency when estimating the coefficients. For simplicity, we ignore this in the rest of the paper. SFA is thereby also a downwards offset of OLS or an upwards offset of COLS.

As we showed in equation (23), the DEA models rely heavily on the error term from OLS regression models (OPEX regression and CAPEX regression). Using the same OLS error term, we can write the formula for the COLS efficiency score. However, in a COLS setup, OPEX and CAPEX should be calculated simultaneously in the same regression model instead of two separate models[16]. Therefore, we only have a single x and a single ϵ, which (in theory[17]) is the sum of the error terms for OPEX and CAPEX, respectively.

[16] We could alternatively calculate two separate COLS models and add the results afterwards if we want the COLS method to follow the DEA method used by DWRA.

[17] In practice, the used solver will most likely find different results for the two methods

$$f_{COLS} = \frac{x^0 - \epsilon_{COLS}^0 + \epsilon_{COLS}^{k^*}}{x^0}, \qquad for\ k^* = peer = \min \epsilon_{COLS}^k \tag{24}$$

Equation (24) calculates the distance to the OLS cost function and offsets the value by the error term for the company with the lowest error (the most efficient company). We divide the formula with the costs to calculate the relative efficiency to compare the result with the DEA efficiency score. Note that ϵ_1^k is always negative. Therefore, the COLS efficiency score is defined as the absolute distance between the error term between the company under evaluation and its peer.

Therefore, the DEA model in (23) and the COLS model in (24) are pretty similar as they both rely on error terms from a standard OLS model. The difference is that we in DEA use relative errors terms and COLS absolute error terms. In addition, DEA weight OPEX and CAPEX with the dual multipliers where COLS does it directly in the OLS model.

In SFA, we need some more assumptions to calculate a similar formula. As previously mentioned, we assume that the coefficients are the same as in OLS. In addition, we assume that all companies have the same noise. These assumptions are, of course, not met in reality, but for the sake of this specific discussion, this does not matter. We write the equation for the SFA efficiency scores in (25) using these assumptions.

$$f_{SFA} = \frac{x^0 - \epsilon_{COLS}^0 + \epsilon_{COLS}^{k^*} + v_{SFA}^0}{x^0}, \qquad for\ k^* = peer = \min \epsilon_{COLS}^k \tag{25}$$

In this SFA setup, we use the same error terms as in COLS but split them into inefficiency and noise. By adding the noise term, v_{SFA}^0, we offset the COLS cost function upwards, as illustrated in Figure 10. We have, therefore, shown that the DEA approach used by the regulator is similar to both COLS and SFA.

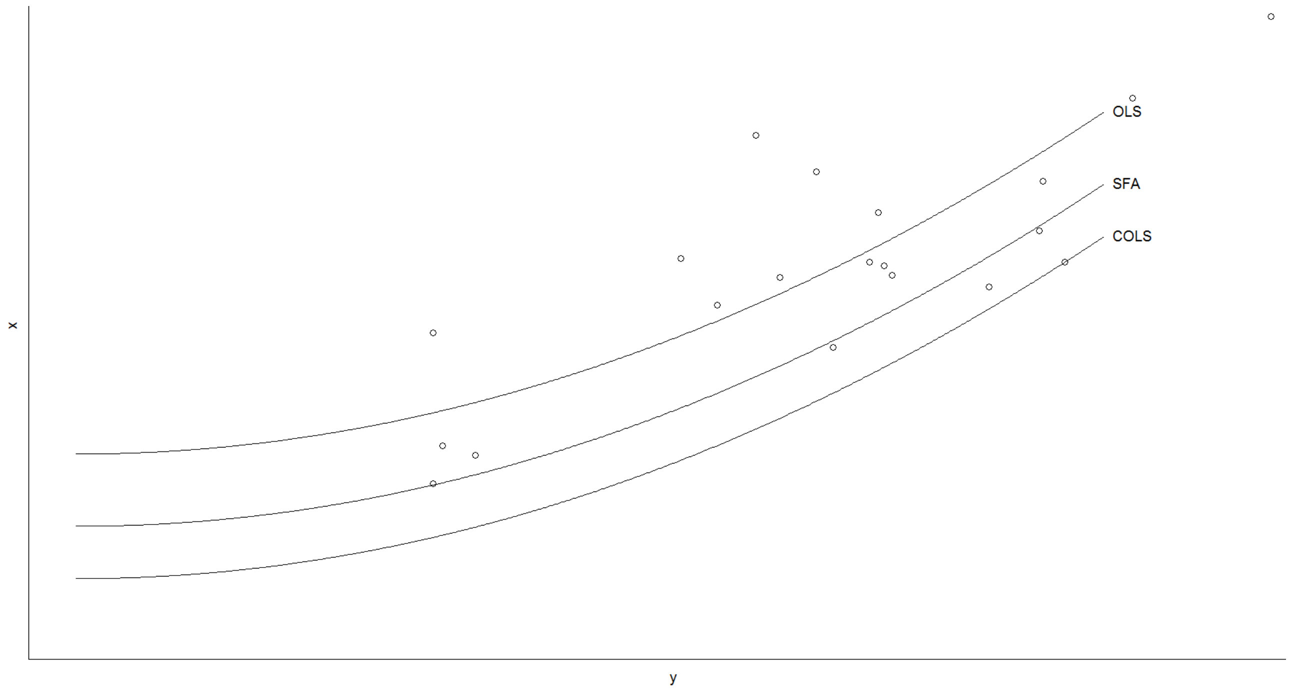

Figure 10 – Illustrative example of the connection between COLS, SFA and OLS (which is used to create the net volumes)

One advantage of using DEA is that it is non-parametric. However, the regulators DEA model relies heavily on parametric assumptions, which means this advantage diminished. At the same time, the model does not get the advantage of the noise term as in SFA. The DEA model is still non-parametric concerning the weight between OPEX and CAPEX, but this seems like the only real advantage left in this simple one-input and two-output model.

However, we have until now assumed that the regulator calculates prices every year. This is not true. In reality, the regulator uses the same prices every year until they decide that the prices are outdated. In this way, they save resources by not doing these calculations and more importantly, the companies save resources, as they only need to present their total costs for the regulator instead of splitting them up to each regression, cf. the discussion around equation (17). If the regulator wants to use the COLS or SFA approach above, they and the companies will, therefore, need to spend extra resources on collecting the underlying data and estimating the models.

6 Conclusion

In DEA, it is common to aggregate inputs and outputs due to the "curse of dimensionality". There are several methods for aggregating the inputs and outputs, each with advantages and disadvantages. A popular method is to use a price-based approach, which corresponds to calculating a combination of technical- and allocative efficiency.

This paper introduced uncertainty into the aggregation method and examined how volatile the efficiency scores are to this uncertainty. We used the Danish waste water sector to show that the average efficiency scores were stable to aggregation uncertainty. However, the results show big changes in the efficiency scores for the individual companies if the aggregation methods contained uncertainty. In the most extreme cases, we showed, by using weight restrictions, that an aggregation uncertainty of 20 % in some cases would change an efficiency score with up to 26 percentage points from its initial value.

In addition, we showed how the impact of aggregation uncertainty is influenced by the companies' (dis)similarities. The more the companies differ, the more critical it is to reduce the uncertainty. In the Danish waste water sector, most companies are pretty similar. However, a few companies have a different output mix than the average companies, which means that they are highly volatile to aggregation uncertainty.

Lastly, we showed how a specific aggregation scheme used by the Danish waste water regulator and many other European utility regulators converts the traditional non-parametric DEA model to a semi-parametric model with close ties to the COLS and SFA methods. We argue that these DEA models lose one of their main advantages, namely that they are non-parametric.

Regulators and other DEA practitioners should closely consider how they aggregate inputs and outputs and how considerable the corresponding uncertainty is. The Danish waste water regulator should consider how big an uncertainty they are willing to transfer to the companies and maybe find an alternative model without as many aggregations as they currently use.

For future research, we suggest comparing the utility regulators current models with simplified versions, where there is no need for these aggregations of hundreds of underlying outputs. If a change in the prices does not change the results much, perhaps there is no need to collect all this data with all its corresponding uncertainties.

In addition, we suggest examining non-linear aggregations schemes in DEA. In this paper, we used linear aggregations in different settings to show how this influences the efficiency scores. In reality, some non-linear aggregations might be better suited to describe the underlying cost-drivers to, for example, take into account different returns to scale. It is easy to aggregate the underlying cost-drivers prior to DEA using non-linear aggregations. It is, however, more complicated to do this in a setup where the aggregations are created with weight restrictions, as this probably will require non-linear optimization programs.

Finally, we suggest more research regarding significant tests between two models. This paper proposes to adjust the methods developed in Asmild, Kronborg, & Rønn-Nielsen (2018) and Rønn-Nielsen, Kronborg, & Asmild (2019) but has not considered the underlying statistical assumptions and properties.

7 Bibliography

Agrell, P. J., & Bogetoft, P. (2017). Regulatory Benchmarking: Models, Analyses and Applications. *Data Envelopment Analysis Journal*(3).

Agrell, P., & Bogetoft, P. (2018). Theory, Techniques and Applications of Regulatory Benchmarking. In E. Grifell-Tatjé, C. K. Lovell, & R. C. Sickles, *The Oxford Handbook of Productivity Analysis* (pp. 523-555). Oxford University Press.

Aigner, D. J., Lovell, C. A., & Schmidt, O. (1977). Formulation and estimation of stochastic frontier production functions. *journal of Econometrics, 6*, pp. 21-37.

Asmild, M., Kronborg, D., & Rønn-Nielsen, A. (2018, 7). Testing productivity change, frontier shift, and efficiency change. *IFRO Working Paper*.

Banker, R., Førsund, F. R., & Zhang, D. (2017). Use of Data Envelopment Analysis for Incentive Regulation of Electric Distribution Firms. *Data Envelopment Analysis Journal*(3).

Bjørner, T. B., & Jakobsen, J. V. (2021). Price cap regulation and water quality. *Journal of Regulatory Economics, 60*(2), pp. 95-116.

Charnes, A., Cooper, W., & Rhodes, E. (1978). Measuring the efficiency of decision making units. *European Journal of Operational Research*, pp. 429-444.

Dyson, R. G., Allen, R., Camanho, A. S., Podinovski, V. V., Sarrico, C. S., & Shale, E. A. (2001). Pitfalls and protocols in DEA. *European Journal of Operational Research*, pp. 245-259.

Färe, R., & Grosskopf, S. (1985). A nonparametric cost approach to scale efficiency. *The Scandinavian Journal of Economics, 4*, pp. 594-604.

Färe, R., Grosskopf, S., & Zelenyuk, V. (2004). Aggregation bias and its bounds in measuring technical efficiency. *Applied Economics Letters, 10*, pp. 657-660.

Goh, K. H., & See, K. F. (2021). Twenty years of water utility benchmarking: A bibliometric analysis of emerging interest in water research and collaboration. *Journal of Cleaner Production, 284*.

Heesche, E., & Asmild, M. (2020, 3). Controlling for environmental conditions in regulatory benchmarking. *IFRO working papers, 2020*(3).

Heesche, E., & Bogetoft, P. (2021). Incentives in regulatory DEA models with discretionary outputs: The case of Danish water regulation. *IFRO Working Paper, 4*.

Kneip, A., Simar, L., & Wilson, P. (2016). Testing hypotheses in nonparametric models of production. *Journal of Business & Economic Statistics, 3*, pp. 435-456.

Meeusen, W., & Broeck, J. (1977). Efficiency Estimation from Cobb-Douglas Production Functions with Composed Error. *International Economic Review*, pp. 435-444.

Petersen, N. C., & Olesen, O. B. (2015). Facet Analysis in Data Envelopment Analysis. In Z. J., & Springer (Ed.), *Data Envelopment Analysis: A Handbook of Models and Methods* (Vol. 221, pp. 145-190). International Series in Operations Research & Management Science.

Podinovski, V. V. (2004). Production trade-offs and weight restrictions in data envelopment analysis. *Journal of the Operational Research Society, 55*(12), pp. 1311-1322.

Rønn-Nielsen, A., Kronborg, D., & Asmild, M. (2019, 7). Exact tests on returns to scale and comparisons of production frontiers in nonparametric models. *IFRO Working Paper*.

Simar, L., & Zelenyuk, V. (2018). Central limit theorems for aggregate efficiency. *Operations Research, 1*.

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

Zelenyuk, V. (2020). Aggregation of inputs and outputs prior to Data Envelopment Analysis under big data. *European Journal of Operational Research*, pp. 172-187.

A Test for significance in frontier shift and efficiency change

We use a permutation test (Asmild, Kronborg, & Rønn-Nielsen, 2018) to test whether the differences in the models from section 4 are significant. The permutation test is a modification of the one developed for the Malmquist index and, in particular, its decomposition[18] and examines if there is a significant difference between the results from two models.

The test is divided into two parts. First, we test if there is a significant difference between the frontiers in the two models (corresponding to the frontier shift component). Hereafter, we test if there is a significant change in the distribution of the efficiency scores between the two models (corresponding to the efficiency change component).

Following the notation of Asmild, Kronborg, & Rønn-Nielsen (2018), the frontier shift and efficiency change are given in (26)-(27).

$$FS(x, y) = \frac{\hat{\theta}_{t_1}(x, y)}{\hat{\theta}_{t_2}(x, y)} \tag{26}$$

$$EC\left(x_{t_1}, y_{t_1}, x_{t_2}, y_{t_2}\right) = \frac{\hat{\theta}_{t_2}\left(x_{t_2}, y_{t_2}\right)}{\hat{\theta}_{t_1}\left(x_{t_1}, y_{t_1}\right)} \tag{27}$$

FS is the frontier shift, and EC is the efficiency change. $\hat{\theta}$ is an efficiency score, where its subscript t_1 or t_2 indicates the model used to estimate the frontier. Note that where t_1 and t_2 refer to two different periods in a standard Malmquist setup, they are here used to indicate different models. However, the interpretation is similar. x and y denote the input and output. The input is the same in the two models compared here, i.e. $x_{t_1} = x_{t_2}$, whereas the output differs in terms of the aggregation used in the different models.

We use the following procedure:

1) We use the geometric mean of the frontier shift and of the efficiency change between the two models as the test statistic.
2) With probability 0.5 for each company, we switch the inputs and outputs between t_1 and t_2 such that some of the companies keep their original aggregated outputs in both models and the rest switch the aggregation scheme between the models such that the first aggregation scheme is used in the second model and vice versa.
3) We recalculate the geometric mean of the frontier shift and the efficiency change and compare them with the original calculations.
4) We repeat steps 2 and 3 100,000 times to get a statistical distribution.

Suppose the frontier shift and the efficiency change differ significantly between the initial and permutated calculations. In that case, we can conclude that the aggregation uncertainty is significantly essential for the efficiency scores.

[18] Note that the Malmquist index usually compares the same model in two periods. In this paper, we have modified the test to compare two different models.

B Test for significance between aggregations schemes

Rønn-Nielsen, Kronborg, & Asmild (2019) have developed a test based on permutations to analyze whether there is a significant difference between the assumption of CRS and VRS for a given data set. The authors propose to use the geometric mean of the ratio between the efficiency scores from CRS and VRS respectively as a test statistic, cf. equation (28)-(29) below.

$$F_{rts}(x,y) = \frac{\hat{\theta}_{CRS}(x,y)}{\hat{\theta}_{VRS}(x,y)} \tag{28}$$

$$T_{rts} = \prod_{i=1}^{n} F_{rts}(X_i, Y_i)^{\frac{1}{n}} \tag{29}$$

$\hat{\theta}_{CRS}$ is the efficiency scores for CRS and $\hat{\theta}_{VRS}$ is the efficiency scores for VRS. x denotes the inputs, y the outputs and n the number of observations.

Under the assumption of CRS, it is possible to rescale the observations without changing the efficiency scores. If F_{rts} is close to one after rescaling the observations, it is safe to assume CRS. To get a distribution of the test statistic, T_{rts}, the authors propose to use the permutation technique below.

1) Calculate the length of the output vector for each observation, $Z_i = ||Y_i||$ and denote $U_i = \frac{X_i}{Z_i}$ and $V_i = \frac{Y_i}{Z_i}$
2) Permutate the vector Z randomly and denote this $\bar{Z}$
3) Use $\bar{Z}$ to rescale the input and output vectors such that $\bar{X}_i = U_i \cdot \bar{Z}_i$ and $\bar{Y}_i = V_i \cdot \bar{Z}_i$
4) Calculate a new test statistic, T_{rts}^j using the new inputs and outputs
5) Repeat step 2-4 N times, where N is a high number (in this paper $N = 1{,}000$)

The p-value is calculated in (30).

$$\hat{p} = \frac{1}{N}\sum_{j=1}^{N} 1_{\left\{T_{rts}^j \leq T_{rts}\right\}} \tag{30}$$

We modify the method described above to test whether the model with weight restrictions in section 4.3 is significantly different from the initial model used in the Danish water regulation.

First, we reformulate the initial model with OPEX and CAPEX as outputs by splitting up CAPEX into the underlying cost-drivers, multiplying the cost-drivers with their corresponding prices and using these adjusted cost-drivers as outputs together with weight restrictions stating that the output weights need to be equal to each other.

Second, we do the same for the new model with aggregation uncertainty, but instead of using weight restrictions stating that all output weights need to be equal to each other, we allow the weights to vary with $\pm$ 20 %.

By reformulating the models, all outputs now have the same unit of measurement, namely monetary. Therefore, we can substitute one output with another (change the output mix) in the initial model with equal multipliers without changing the results. However, in the second model with aggregation uncertainty, a change in the output mix will yield different results, cf. section 3. We can exploit this to calculate whether the difference between the models is significant, following the principles of the permutation procedure suggested by Rønn-Nielsen, Kronborg, & Asmild (2019) above.

We modify the procedure to the following:

1) Calculate the output mix for each company and denote these vectors M_i
2) Permutate the vectors M_i randomly and denote this $\bar{M}_i$
3) Recalculate the outputs using $\bar{M}_i$ while holding the sum of the outputs of each company fixed:

$$\bar{y}_i = \bar{M}_i \frac{\sum y_i}{\sum \bar{M}_i}, \qquad \forall\, i$$

4) Calculate the test statistic, T^j_{wei} from (28)-(29) using the new outputs and with the relevant model specifications instead of VRS and CRS

The p-value is calculated in (31).

$$\hat{p} = \frac{1}{N}\sum_{j=1}^{N} 1_{\{T^j_{wei} \leq T_{wei}\}} \tag{31}$$