A critical assessment of TCB18 gas

Prepared for
Gasunie transport services B.V.

18 August 2020

Strictly confidential

www.oxera.com

Contents

Oxera Consulting LLP is a limited liability partnership registered in England no. OC392464, registered office: Park Central, 40/41 Park End Street, Oxford OX1 1JD, UK; in Belgium, no. 0651 990 151, branch office: Avenue Louise 81, 1050 Brussels, Belgium; and in Italy, REA no. RM - 1530473, branch office: Via delle Quattro Fontane 15, 00184 Rome, Italy. Oxera Consulting (France) LLP, a French branch, registered office: 60 Avenue Charles de Gaulle, CS 60016, 92573 Neuilly-sur-Seine, France and registered in Nanterre, RCS no. 844 900 407 00025. Oxera Consulting (Netherlands) LLP, a Dutch branch, registered office: Strawinskylaan 3051, 1077 ZX Amsterdam, The Netherlands and registered in Amsterdam, KvK no. 72446218. Oxera Consulting GmbH is registered in Germany, no. HRB 148781 B (Local Court of Charlottenburg), registered office: Rahel-Hirsch-Straße 10, Berlin 10557, Germany.

Boxes, figures and tables

Executive summary

Background

The pan-European Transmission Cost Benchmarking Project[1] (TCB18) was carried out by the Council of European Energy Regulators (CEER) through its consultant, Sumicsid. It was a follow-up to the e2gas study, that was also undertaken by Sumicsid on behalf of the CEER.[2]

The TCB18 gas study ('TCB18 gas') covered 29 TSOs from 13 European countries, including the 16 German gas TSOs, which were not actively involved in the project and whose data was recorded in a different year from the rest of the sample. Sumicsid used cost and asset data provided by TSOs (and in the case of the German TSOs, by their regulator, the Bundesnetzagentur), in addition to environmental and price-level data from external sources, to assess the TSOs' relative efficiency. As in previous benchmarking exercises, Sumicsid used data envelopment analysis (DEA). Unlike in the previous iteration, e2gas, stochastic frontier analysis (SFA) was not considered as a robustness check to validate the model specification and the DEA results.

Gasunie Transport Services B.V. (GTS) has commissioned Oxera to validate and review the results from TCB18 gas, and to recommend robust solutions to any issues that emerge.

Our assessment of TCB18 gas builds on the comprehensive conceptual and empirical analysis we undertook on behalf of all the electricity TSOs that participated in the TCB18 electricity study ('TCB18 electricity'),[3] and on a review of all documents published in TCB18 gas. These include:

- the final report and appendices as published by Sumicsid, which are available on the CEER's website;[4]
- the TSO-specific report for GTS (and TSO-specific reports for other TSOs, where they are publicly available) that summarise individual TSOs' data and performance;[5]
- slides accompanying the five workshops that Sumicsid shared through the course of the TCB18 project;
- the dynamic efficiency report that was published at a later date.[6]

In addition to these documents, we also presented the findings of our assessment of TCB18 electricity in a meeting on 5 June 2020 attended by the ACM, Sumicsid, TenneT B.V. (TenneT), GTS and market parties. While the official transcript of this meeting has not been published, based on our notes, we reference the discussions from this meeting where Sumicsid has provided additional information on its analysis.

[1] Sumicsid (2019), 'Pan-European cost-efficiency benchmark for electricity transmission system operators main report', July.
[2] Sumicsid and Swiss Economics (2016), 'Project E2GAS Benchmarking European Gas Transmission System Operators', June.
[3] Oxera (2020), 'A critical assessment of TCB18 electricity', April.
[4] Sumicsid (2019), 'Pan-European cost-efficiency benchmark for gas transmission system operators main report', July; and Sumicsid (2019), 'Pan-European cost-efficiency benchmark for gas transmission system operators appendix', July.
[5] Sumicsid (2019), 'Project TCB18 Individual Benchmarking Report GTS – 209', July.
[6] Sumicsid (2020), 'Project CEER-TCB18 Dynamic efficiency and productivity changes for gas transmission system operators MAIN REPORT', April.

While we have not had access to the full dataset used by Sumicsid in TCB18 gas (as we had for our assessment of TCB18 electricity), given that the methodologies followed are very similar between the two studies, we have been able to support the validity of our critique using the published documents of the TCB18 study and data on GTS that was used by Sumicsid in its analysis.

GTS has asked Oxera to assess whether the issues raised in our assessment of TCB18 electricity are also relevant for TCB18 gas, and whether there are any additional issues arising from the gas study.

Summary of our assessment

International benchmarking **can be a powerful tool** for companies and regulators to assess the efficiency of network operators. This is especially true in the context of the gas transmission industry, where the sector is often characterised by national monopolies, thus making national benchmarking challenging. In this sense, we welcome projects such as TCB18 and its predecessors (e.g. e2gas), which have attempted to develop a framework for periodic assessment of TSOs.

Nevertheless, the TCB18 study itself suffers from a **number of significant flaws**, some of which are **fundamental**. These flaws mean that the estimated efficiency scores and suggested cost savings are not robust and **thus cannot be used in their current form for regulatory purposes** (or operational or valuation purposes). While some of these flaws simply add to the general uncertainty in the estimated efficiency scores for all TSOs, some flaws (such as the adjustment for price levels and the failure to account for heterogeneity such as density) may bias the TCB18 gas study against GTS.

Some issues, like consistency in reporting guidelines, are partly driven by the lack of maturity in the international benchmarking process, and we expect this to improve with time. This is particularly true in gas transmission benchmarking, which is less developed than electricity transmission benchmarking. However, Sumicsid's concluding remarks are concerning, as they are not consistent with the significant issues and areas for future work identified through our comprehensive review. For example:

> Regulatory benchmarking has reached a certain maturity through this process and model development, signaling both procedural and numerical robustness […]
>
> […] future work can be directed towards further refinement of the activity scope and the interpretation of the results, rather than on the model development.[7]

As demonstrated throughout this report, Sumicsid has not ensured that its dataset is sufficiently free of data errors, has not developed a model that adequately accounts for all material drivers of expenditure, and has not appropriately considered valid alternatives to its estimation and validation procedures. Indeed, in our review of TCB18 electricity, we presented several sources of evidence that indicated that much or all of the estimated efficiency gaps in electricity can be explained by model misspecification and statistical noise. Moreover, in additional work for TenneT, we found specific examples of where Sumicsid's approach leads to a material bias for individual TSOs, on top of the general uncertainty. As such, future iterations of the study require significantly more changes than simple refinements to the activity scope. If the

[7] Sumicsid (2019), 'Pan-European cost-efficiency benchmark for gas transmission system operators', July, p. 49.

recommendations presented in this report are considered in future iterations of the study, we consider that the CEER will be better able to develop a process and methodology for international cost benchmarking that are informative and fit for purpose.

Key themes of our assessment

The main issues with Sumicsid's analysis are categorised under four themes, as follows.

1. The TCB18 process and methodology was not sufficiently transparent to allow third parties to follow or validate Sumicsid's analysis.

An overarching theme of the TCB18 project, covering both the gas and electricity studies, is the lack of transparency. That is, Sumicsid's outputs do not contain the necessary information for third parties (including the participating TSOs themselves) to follow its analysis, validate its analysis or its sources. In our review of TCB18 electricity, we identified a number of Sumicsid's statements that we could not validate, despite having access to the full dataset used in TCB18 electricity. This indicates that either: (i) Sumicsid's description of its approach is not sufficient to allow third parties to replicate its analysis; or (ii) Sumicsid has made statements that are incorrect.

Not only is transparency important for third parties to follow the analysis and understand the limitations of the project's conclusions, but there are potentially valuable insights from DEA modelling that can aid TSOs' (and their national regulatory authorities, NRAs') understanding of their (the TSOs') efficiency. For example, since DEA determines a TSO's efficiency by comparing that TSO with (a combination of) efficient TSOs that are supposed to operate under structurally comparable conditions (its 'peers'), knowing which TSOs are peers could help TSOs that are estimated to be inefficient to improve their efficiency by adopting best practice. This can be done through knowledge transfer across TSOs and a deeper understanding of their operations. Relevant outputs to enable such learning were not discussed in TCB18.

Thus, the level of transparency exhibited by Sumicsid falls short of what is commonly considered good practice in regulatory contexts.

2. Sumicsid's data collection and construction processes do not enable a sufficiently harmonised dataset to undertake robust cost benchmarking.

Sumicsid has not adequately ensured that the final dataset is free from significant data errors and inconsistencies.

Sumicsid has stated that it carried out a rigorous data-collection exercise involving a number of iterations with the TSOs and NRAs.[8] In theory, the procedure followed should have produced a relatively robust dataset for benchmarking purposes. However, there is strong *prima facie* evidence that several material issues remain. For instance:

[8] Sumicsid (2019), 'Pan-European cost-efficiency benchmark for gas transmission system operators main report', July, section 3.2.

- **The process described may lead to biased outcomes**. One stage of the data validation is a check of the data by the NRA. It is not clear if and to what degree the NRAs followed common guidelines or proceeded with the same level of care in scrutinising the data. For instance, the ACM decided to exclude some of GTS's connection points from the data. If this was not done by other NRAs, the data is biased against GTS. Furthermore, we understand from GTS that the level of engagement varied widely across participating TSOs and there has not been a discussion with all participating TSOs concerning the costs (and cost definitions), activities and assets to be included in the dataset. This adds another source of uncertainty in the data-collection process.
- **The process described might not have been followed rigorously**. In our review of TCB18 electricity, we found cases of clearly erroneous data that was in violation of Sumicsid's stated validation criteria. For example, Sumicsid's criteria for data validation in electricity stated that the proportion of angular towers should be at least 10%,[9] yet one TSO in the electricity study is reported to have no angular towers. Moreover, Sumicsid's calculation of transformer power is biased against TSOs that did not own some of the transformers connected to their grids, and alternative measures of network capacity (such as circuit ends power) do not appear to have been considered by Sumicsid.[10] As the data collection and validation exercises are largely the same for both TCB18 gas and TCB18 electricity, it is possible (or likely) that similar data issues arise in the gas study.
- **The same process was not followed for the German gas TSOs, which constituted the majority of the sample, as for other TSOs**. The German TSOs were not part of TCB18 apart from their attendance of the initial workshop and, therefore, we understand that they could not ensure the accuracy of their data. Moreover, Sumicsid did not have some of the data needed and thus had to make assumptions that introduced additional data uncertainty. For instance, the power of compressors is an input in calculating NormGrid, but this information was not available for the German TSOs.[11]

Given the lack of involvement of the German gas TSOs, and the material data errors we identified in our review of TCB18 electricity (which followed a similar data-collection process), it is likely that data errors are present in TCB18 gas. Indeed, we understand that several data errors do exist in the gas dataset, including the allocation of planning costs, the allocation of indirect costs and the calculation of capital expenditure (CAPEX) for certain assets.[12] Where data errors exist, the best course of action is to correct them directly, rather than relying on outlier tests to isolate suspicious firms.

While difficulties may exist in correcting some of the data errors, the issue must not be ignored—instead, some form of adjustment and sensitivity analysis to account for data/modelling errors is required. Indeed, as Sumicsid's application of DEA is deterministic, a TSO's estimated inefficiency could be driven by statistical noise. Adjustments for statistical noise could take several forms. A

[9] This was stated by Sumicsid in a validation document for one of the electricity TSOs. See Sumicsid (2018), 'TCB18 Crossvalidation TECH/ELEC', October, p. 1.
[10] Oxera (2020), 'Analysis on TenneT's estimated efficiency under TCB18', August, section 2.
[11] Sumicsid (2019), 'Pan-European cost-efficiency benchmark for gas transmission system operators main report', July, p. 32.
[12] In the construction of CAPEX, Sumicsid has dismissed the issue of significant rehabilitations (i.e. large investments to extend the life of existing assets) on the grounds that there is too much uncertainty in the adjustment. However, ignoring the issue also introduces uncertainty in the data, and this uncertainty (alongside all other forms of uncertainty) should be acknowledged in Sumicsid's outputs.

simple approach would be to adjust the efficiency score of each TSO such that it is compared with a less stringent benchmark than the frontier. It is important to note that the benchmark cannot be arbitrarily determined and should be supported by empirical evidence regarding the data and model quality and the confidence in the results.

It is also possible to cross-check the results from DEA with other methods, such as SFA, that are able to account for certain types of data error.[13] Such cross-checking is common in German energy regulation. In our review of TCB18 electricity, we found that **the 'efficiency gap' estimated by Sumicsid could be entirely attributed to unmodelled heterogeneity or statistical noise**, thus indicating that caution must be exercised when interpreting Sumicsid's analysis. Indeed, this finding supports the conclusion that the estimated efficiency scores cannot simply be interpreted as differences in genuine efficiency across TSOs.

Another possibility is to simulate the impact of data uncertainty on the DEA model through Monte Carlo analysis. In our review of TCB18 electricity, we found that adjustments of 10–18 percentage points to the estimated efficiency scores could be appropriate based on uncertainty in the electricity dataset.[14] In the case of TCB18 gas, given the additional layer of uncertainty caused by the inclusion of German TSOs' data, an adjustment towards the top end of this range (or above the top end of the range) may be more appropriate, and will need to be empirically determined.

In general, for the reasons described above, the quality of the data is not likely to be as good as is being presented by Sumicsid in its report. This should be taken into account when interpreting the efficiency scores from the DEA model.

Sumicsid's choice of input variable does not appropriately capture the trade-off between different types of expenditure.

Sumicsid has modelled expenditure on a total expenditure (TOTEX) basis, where TOTEX is the sum of operating expenditure (OPEX) and CAPEX. This implicitly assumes that there is a one-to-one trade-off between OPEX and CAPEX. However, OPEX and CAPEX are calculated differently (and are not strictly a measure of a TSO's expenditure in a given year) and subject to different normalisations[15] that may limit the extent to which the two types of expenditure are comparable. If TSOs have different ratios of OPEX to CAPEX dictated by national regulatory and legislative frameworks and operational characteristics,[16] TOTEX modelling, as considered by Sumicsid, could inappropriately conflate TSO heterogeneity that is beyond management control with inefficiency.

Moreover, GTS's estimated efficiency is highly sensitive to the value of the weighted average cost of capital (WACC) assumed in the modelling to derive CAPEX. Its estimated efficiency increases from 73.2% to 81.4% when the WACC parameter reduces from 3% to 1.8%,[17] the latter being more aligned

[13] Specifically, basic applications of SFA assume that statistical noise (i.e. data errors and modelling uncertainty) is independent and identically distributed (i.i.d) and follows a normal distribution.
[14] Oxera (2020), 'A critical assessment of TCB18 electricity', April, section 3.1.
[15] For example, OPEX is calculated on an annual basis and is adjusted for differences in labour input prices. CAPEX, on the other hand, is calculated as the sum of annuities in specific investments, and no adjustment is made for differences in input prices.
[16] For example, in our review of TCB18 electricity, we found that one TSO did not own the assets it operates. Therefore, its CAPEX (as a share of TOTEX) was significantly lower than the rest of the sample. Indeed, the share of CAPEX within TOTEX varied from 23% to 90% across the sample. See Oxera (2020), 'A critical assessment of TCB18 electricity', April, section 3.2.
[17] Sumicsid (2019), 'Project TCB18 Individual Benchmarking Report GTS – 209', July, Figure 4.9.

with the WACC parameter used by the ACM to set allowed revenues. This could indicate that GTS is being compared with TSOs that are less CAPEX-intensive than GTS, some of which may not be appropriate comparators due to different operational or legislative characteristics.

There are several ways to account for this heterogeneity, none of which have been properly examined by Sumicsid. For example, OPEX and CAPEX can be kept as separate inputs in the DEA model; this allows a broader range of trade-offs and ensures that TSOs are only benchmarked against peers with comparable levels of OPEX to CAPEX, and mitigates the risk that a TSO is benchmarked against an incomparable peer with a very different cost structure. Alternative approaches include developing separate models (econometrically or through DEA) for OPEX and CAPEX, while recognising the trade-offs between the two and without imposing unnecessary assumptions.

Sumicsid has not sufficiently accounted for differences in input prices across TSOs.

Sumicsid adjusts only for direct manpower OPEX, using an index of civil engineering price levels to account for differences in input prices across TSOs. This translates to just 6% of GTS's TOTEX being normalised for price-level differences. No adjustment is made to CAPEX or other cost items within OPEX. This approach raises a number of issues, each of which likely affects GTS's estimated efficiency.

- The civil engineering price-level index (PLI) contains prices for non-labour inputs (such as raw materials like metals, plastics and concrete). Its application to labour costs is therefore insufficiently substantiated. Actual PLIs for labour costs were dismissed without sufficient justification.
- Sumicsid has limited the scope of the adjustment to a specific cost line within OPEX. In reality, 40–60% of CAPEX is driven by labour or labour-related costs according to estimates from Sumicsid and GTS.
- The differences in non-labour input prices, such as machinery and services (which would have an impact on both OPEX and CAPEX), are not directly accounted for at all.
- Much of the (44-year) CAPEX investment stream used to calculate TOTEX was incurred decades ago, before the close integration of the EU and the EEA was complete.[18] As such, Sumicsid's implicit assumption that TSOs operated in the same market when historic investments were undertaken (and can therefore purchase inputs at the same prices) is not sufficiently motivated.

On adjustment for input prices to labour costs, Sumicsid must consider a relevant PLI for labour costs, rather than an a PLI that includes many factors of production, including raw materials (as it has done in this case). Furthermore, if labour costs are to be adjusted in OPEX, they should also be adjusted in CAPEX. Using data on labour cost components in CAPEX provided by GTS, we estimated adjustments to CAPEX based on the proportion of CAPEX which is related to labour and a PLI (or labour cost index) reflecting labour cost differences, as has been considered in other international benchmarking applications. Price levels can vary across participating countries by up to 80%. Thus, deciding which index to use and how much of the cost base should be

[18] For instance, GTS incurred nearly 50% of its CAPEX in TCB18 prior to 1990 and thus well before the free movement of labour and capital within the EU was introduced.

adjusted will have a *significant* impact on the estimated efficiency scores. Indeed, in our review of TCB18 electricity, we found that the estimated efficiency of nearly all TSOs were highly sensitive to the choice of PLI.[19]

In our review of TCB18 electricity, we noted that an alternative option would be to adjust *all* OPEX with the PLI for overall GDP and to adjust *all* CAPEX with the PLI for civil engineering, as the activities contained in these indices appear aligned with TSOs' activities. In this way, differences in prices for all relevant inputs can be accounted for, rather than Sumicsid's focus on labour-cost differentials.

While GTS is not uniquely affected by this issue, the Netherlands is a relatively high price-level country (compared with the rest of the TCB18 sample). Therefore, it is likely that at least part of the estimated efficiency gap for GTS is a result of uncontrollable relative price levels.

Sumicsid's allocation of indirect costs to assessed OPEX is arbitrary, and evidence supporting its allocation rule is not presented in the report.

Sumicsid has allocated indirect costs (e.g. human resources expenditure, IT support) to activities considered within the scope of benchmarking based on a unsubstantiated allocation rule. Specifically, it has allocated indirect expenditure to activities based on the percentage of OPEX (minus energy costs and depreciation) in that activity. Large, uncontrollable cost items that are unrelated to indirect expenditure (such as taxes and levies) can have a significant impact on the amount of expenditure allocated to in-scope activities.

We recommend amending the allocation rule to exclude all costs that are considered outside the scope of benchmarking. This mitigates the risk that indirect expenditure is arbitrarily allocated to activities based on cost items that are unrelated to indirect expenditure. The allocation of indirect expenditure is an important conceptual issue that can have a material impact on the TSOs' estimated efficiency.

3. Sumicsid's approach to model development appears arbitrarily restrictive and inconsistent with the scientific literature.

A robust model-development process is necessary to ensure that the results from an empirical investigation are unbiased and robust to challenge. This process should take into account the operational and economic rationale for including or excluding specific cost drivers and should be supported by both statistical analysis and operational evidence. Sumicsid's model-development process is not clearly presented in any of its outputs, nor does it consistently follow scientific best practice.

Sumicsid's cost driver analysis is not transparent and contradicts operational intuition.

Sumicsid has not presented a coherent model-development process which led to the final model being selected. Such a process should involve: (i) the selection of candidate cost drivers based on operational insight and interactions with the industry; (ii) detailed statistical analysis of the candidate cost drivers to identify the most appropriate to include in the final (set of)

[19] Oxera (2020), 'A critical assessment of TCB18 electricity', April, Figure 3.3.

model(s); and (iii) robust sensitivity analysis with respect to the choice of cost drivers and modelling assumptions.

It appears that Sumicsid used a combination of ordinary least squares (OLS) regression (with and without outliers) and 'robust OLS' (ROLS) regression to calculate model fit. The model size appears to be based on Lasso regression. Sumicsid has not presented analysis behind its model-development process in its final reports,[20] but it has presented the coefficients of an ROLS regression on its final model in the TSO-specific outputs.[21] What is visible of the model development is flawed for the following reasons.

- The econometric results are in violation of operational intuition. In Sumicsid's final benchmarking model, two of the four cost drivers used, which should have a positive impact on costs, have an estimated *negative* sign.[22] This is counterintuitive and not aligned with operational rationale. Indeed, Sumicsid itself has stated the need for cost drivers to be positively associated with costs explicitly in previous benchmarking studies.[23]
- ROLS is not an appropriate standalone tool in model development. This estimator explicitly gives less weight to observations that are further from the regression line, which is assumed on the dataset. Instead, Sumicsid should have tried to understand why these TSOs did not fit the model well, e.g. did they face additional cost drivers or have a different operating model?
- R-squared is not the sole informative measure of model quality. R-squared is entirely based on model fit; it does not, for example, take into account whether the model is supported by operational intuition (which was clearly violated in TCB18 gas) and thus must always be supported by an examination of the regression results. For example, R-squared is indifferent between the positive and negative coefficients in the model. In addition, Sumicsid appears to have inflated this metric, such as by suppressing the intercept.
- Lasso regression cannot be used as evidence to make a *general* observation about optimum model size for a dataset. Rather, an appropriate application of Lasso (which would have to be accompanied and validated by other evidence) could work with a *specific* set of cost drivers and selects a *specific* set of drivers as being the most important ones based on a specified validation criterion.

As a result of these modelling flaws, it is unlikely that Sumicsid's model-development process has led to an appropriate final model from which unbiased efficiency scores could be estimated. For an appropriate model to be identified, Sumicsid should clearly outline its process and candidate cost drivers to the project participants. Participants need to have the opportunity to critique the approach and the candidate cost drivers, and should be allowed to suggest alternatives for both for consideration before arriving at the final (set of) model(s).

[20] Some alternative models are presented in workshops throughout the TCB18 study. However, the final model was not justified in these workshops.
[21] Sumicsid (2019), 'Project TCB18 Individual Benchmarking Report GTS – 209', July, Table 3.1.
[22] Sumicsid (2019), 'Project TCB18 - Individual Benchmarking Report - GTS – 209', July, Table 3.1.
[23] Sumicsid, Frontier Economics, and Consentec (2013), 'E3GRID2012-European TSO Benchmarking Study', July, p. 130.

Sumicsid's use of NormGrid is inconsistent with the non-parametric nature of DEA and does not account for modelling uncertainty.

Sumicsid considers NormGrid to be 'the strongest candidate in the frontier models'.[24] Constructed variables such as NormGrid carry an inherent risk of favouring some TSOs at the expense of others; this is because such variables reflect an aggregation of asset characteristics using assumed formulas that are subsequently further aggregated across a number of asset classes using weights that are themselves estimated with a degree of uncertainty. This issue is particularly relevant if there is significant heterogeneity in terms of the assets that TSOs deploy to meet their service obligations, as is the case in TCB18 gas.[25]

The aggregation of asset characteristics involves assumptions about the cost impact of these characteristics. This is explicitly stated by Sumicsid at various points, but no sensitivities or margins of error are presented. The validation exercise for pipelines on ACER data that Sumicsid considered shows that there is considerable variation in costs when controlling for Sumicsid's parameters.[26] The 'noise' in these estimates and assumptions made are not addressed at any point in Sumicsid's analysis.

The aggregation across asset classes adds further uncertainty. Depending on the weights used and the mix of assets of each type that a TSO uses in reality, a TSO may be favoured or disadvantaged relative to other TSOs that differ substantially on the mix of assets. The weights used in the aggregation of NormGrid should therefore be robustly justified.

Sumicsid has stated that it used linear regression to derive the appropriate OPEX and CAPEX weights on NormGrid.[27] The results of this regression analysis are not presented in any of Sumicsid's outputs. We were unable to validate the NormGrid weights in our review of TCB18 electricity and, given the lack of transparency in Sumicsid's outputs, the gas study suffers from the validation issue as well. We can assume that the weights cannot be validated in TCB18 gas.

Deriving weights based on regression analysis relies on parametric assumptions that are inconsistent with the non-parametric nature of DEA. In this context, it may be more appropriate to consider each asset class as a separate output and to allow the DEA model (if required with weight restrictions) to determine the correct weights on each asset class. If Sumicsid is concerned that a TSO could place an excessive weight on one particular asset class, it can introduce weight restrictions. A weight-restricted DEA model would offer more flexibility than Sumicsid's rigid assumptions regarding the importance of each asset.

The environmental adjustments are arbitrary, inappropriate and overly restrictive.

Sumicsid has chosen to model environmental factors by multiplying certain cost drivers with 'complexity factors'. This approach and its application in

[24] Sumicsid (2019), 'Pan-European cost-efficiency benchmark for gas transmission system operators main report', July, p. 33.
[25] For example, we understand that the gas TSOs in the sample can be split into three general groups: (i) transit TSOs that only transport gas across large distances and therefore have a disproportionate amount of pipelines; (ii) distribution TSOs that do not transport gas over large distances thereby being more akin to gas DSOs and having a disproportionate amount of compressors; and (iii) mixed TSOs that fulfil both functions. GTS is a mixed TSO.
[26] Sumicsid (2019), 'Norm Grid Development - TCB18 PROJECT', February, Figure 2-3.
[27] Sumicsid (2019), 'Norm Grid Development - TCB18 PROJECT', February, p. 2.

TCB18 are not sufficient in addressing environmental complexity, for the following reasons.

- Sumicsid does not appear to have undertaken simple validation exercises to ensure that its environmental adjustment is appropriate on the data. For example, in our assessment of TCB18 electricity, we found that the environmental complexity factor was *negatively* correlated with unit costs, meaning that TSOs operating in more complex regions have lower costs. This has no operational rationale. Given that such an exercise was not performed for TCB18 electricity, and Sumicsid has presented no evidence to suggest that it was performed for TCB18 gas, we consider it unlikely that this essential validation exercise has been performed.
- Sumicsid has not ensured that all relevant environmental factors are adequately accounted for in its model. Sumicsid argues that correlations between factors are sufficient, but some factors that Sumicsid has omitted from its final model (e.g. land use) are in fact negatively correlated with the slope factor used.
- Sumicsid's approach is equivalent to imposing strict weight restrictions on environmental complexity that it has not justified with empirical evidence. For example, it is not clear why slope complexity should affect a TSO only if weight is placed on NormGrid when estimating its efficiency score. If a TSO attaches no weight to the NormGrid variable when estimating its efficiency score, it will not benefit from the adjustment, regardless of how complex its operating environment is.
- In addition, Sumicsid has not justified why it has adjusted two output variables (NormGrid and pipe length) by environmental factors (by slope and humidity, respectively), and not the other two outputs (compressor power and connection points).
- Sumicsid's environmental adjustment factors ignore asset location. The complexity factors appear to be calculated on a country level. If assets are not evenly distributed across the country and are, for example, more likely to be located in flat, densely populated areas, this will create a disconnect between the costs and cost drivers and thus bias results. Indeed, we understand that GTS has provided evidence that the environmental adjustment can be overstated by up to 50% for some TSOs.[28]

Importantly, Sumicsid has presented no evidence to suggest that the impact of population density on costs is robustly accounted for in its model. As population density is a well-established driver of gas transmission costs,[29] and GTS operates in one of the densest environments in the sample, this may result in a significant bias in GTS's estimated efficiency score. Indeed, in our analysis of TenneT's (the Dutch electricity TSO) efficiency, we found that controlling for density in the model can increase its estimated efficiency from 71.5% to 100%.[30]

4. Sumicsid has not justified the assumptions that it has made in its model, and its approach to model validation is incapable of detecting flaws or omissions in its model.

[28] Based on GTS analysis of GIS data for [*Vertrouwelijk: a European country*] .
[29] Indeed, Sumicsid itself notes that density can be a driver of costs for pipelines. See Sumicsid (2019), 'Norm Grid Development', February, p. 22.
[30] Oxera (2020), 'Analysis of TenneT's estimated efficiency under TCB18', August, section 4.

Sumicsid has arbitrarily excluded other valid benchmarking approaches.

DEA is one method of assessing the efficiency of TSOs. Although it has certain advantages over other methods, it is a deterministic method of efficiency assessment (as applied by Sumicsid). It is therefore common for regulators and consultants to use alternative benchmarking methods, either to directly inform the efficiency target after accounting for statistical noise or as a cross-check to the results from DEA, as Sumicsid did in e2gas when it used SFA to cross-check results.

Sumicsid has stated that it has not used SFA in this case as 'the number of observations is too small'.[31] However, Sumicsid has access to data for 29 TSOs and 70 observations (when data over time is accounted for)—this is significantly bigger than the dataset Sumicsid used in e2gas where SFA was applied, and is bigger than other datasets where SFA has been used in some regulatory applications.[32] Moreover, Sumicsid uses other econometric tools (e.g. OLS, ROLS, Lasso) in the model development and model validation that also have similar data requirements and perform better on larger samples.

The appropriateness of a method has to be determined empirically on the data and model used. The size of the dataset can therefore not be considered a valid reason for excluding SFA in TCB18 gas. SFA can offer valuable insight, especially if the data is prone to errors, as is the case in TCB18.

Sumicsid has provided no evidence to support its returns to scale assumption and dismissed the use of alternative assumptions without valid reason.

Sumicsid has stated that the non-decreasing returns to scale (NDRS) assumption is supported by statistical evidence, but has not presented the evidence (unlike in e2gas, where Sumicsid presented the results from one statistical test to support its returns to scale assumption).[33] In our critique of TCB18 electricity, we performed the same tests as Sumicsid and found no conclusive evidence supporting an NDRS assumption.[34] Indeed, some tests supported a variable returns to scale (VRS) assumption, counter to Sumicsid's statements.

At the meeting with the ACM on 5 June 2020, Sumicsid stated that its returns to scale assumption was not primarily determined by statistical tests but was based on conceptual reasoning and the academic literature. This casts doubt on the validity of Sumicsid's statistical statements in its main report.[35] Given that we found not statistical evidence to support the returns to scale assumption in TCB18 electricity, it may also be the case that the statistical evidence does not support Sumicsid's returns to scale assumption in TCB18 gas. In this case, if the observed relationship between cost and output is not consistent with Sumicsid's (unjustified) expectations, Sumicsid should have investigated and documented the causes of such inconsistencies instead of ignoring the problem. This could have led to the detection of omitted cost

[31] Sumicsid (2019), 'Pan-European cost-efficiency benchmark for gas transmission system operators main report', July, p. 30.
[32] For example, the Office of Rail and Road (ORR) performed SFA on a sample of 50 observations for its determination of the efficiency of Network Rail as part of the PR18 price control. Office of Rail and Road (2018), 'PR18 Econometric top-down benchmarking of Network Rail A report', July.
[33] Sumicsid and Swiss Economics (2016), 'Project E2GAS Benchmarking European Gas Transmission System Operators', June, p. 43.
[34] Oxera (2020), 'A critical assessment of TCB18 electricity', prepared on behalf of European electricity TSOs, April, section 5.1.3.
[35] Sumicsid (2019), 'Pan-European cost-efficiency benchmark for gas transmission system operators main report', July, section 5.12.

drivers, such as urbanity (two of the largest TSOs in the TCB18 sample, GTS and National Grid, operate in some of the most densely populated countries in the sample) or asset age, which could be driving the estimated inefficiency of large TSOs.

If the technology is indeed VRS, there is no reason for the technology to be restricted in an international benchmarking context. As the scale is largely dictated by national boundaries, it is unlikely that re-organisation could ever yield TSOs with a cost driver mix of 'optimum scale'. Aside from the practical difficulties associated with such a re-organisation, some TSOs (including GTS), are legally prohibited from separating into two or more TSOs.

Sumicsid's scale efficiency analysis implies that GTS would be estimated to be fully efficient under a VRS assumption.[36] Given the sensitivity of GTS's efficiency to this assumption, GTS is being compared to TSOs that may not be operationally comparable to GTS. For example, the solutions available to smaller TSOs in the peer group may not be feasible for GTS.

Sumicsid's outlier procedure is not justified in its report and is scientifically inadequate.

Sumicsid has relied on German regulatory precedent to detect outliers. Specifically, Sumicsid has performed dominance and super-efficiency tests based on the Bundesnetzagentur's approach to outlier detection. The decision to follow the outlier procedure specified in the ARegV is not justified, nor is the outlier procedure likely to be sufficient in an international benchmarking context.

We recommend the following amendments to Sumicsid's outlier procedure.

- **Dominance test**. Following the recommendations of Kumbhakar, Parthasarathy and Thanassoulis (2018)[37] in their expert opinion on Sumicsid's dominance test, we would recommend applying a 'bootstrap-based test' for dominant TSOs. Sumicsid's dominance test has no theoretical foundation in the context of outlier analysis and is typically biased against the detection of outliers. The bootstrap-based test provides a robust foundation as it is a non-parametric test consistent with DEA and can better take into account the specific context (i.e. outlier analysis).
- **Super-efficiency test**. Sumicsid's application of the super-efficiency test is unable to detect 'masked' outliers[38] and is less capable of detecting outliers in a volatile sample. Consistent with the recommendations in Deuchert and Parthasarathy (2019)[39] and Thanassoulis (1999),[40] we would recommend setting an absolute threshold for the detection of outliers and applying the

[36] GTS is estimated to be 73.2% efficient under NDRS by Sumicsid. The scale efficiency, defined as $\frac{DEA_{CRS}}{DEA_{VRS}}$ is also estimated to be 73.2%. Given that GTS is the second largest TSO in the sample, the NDRS and CRS efficiencies are likely to be the same. If this is the case then the efficiency under VRS would be 100%. Note that this observation assumes that there is no interaction between the returns to scale assumption and the outliers detected under the different assumptions.
[37] Given the non-applicability of the ARegV in the current context, as noted in Kumbhakar, Parthasarathy and Thanassoulis (2019), the test can be easily amended to improve on its discriminatory power. See Kumbhakar, S., Parthasarathy, S. and Thanassoulis, E. (2018), 'Validity of Bundesnetzagentur's dominance test for outlier analysis under Data Envelopment Analysis', August.
[38] Masked outliers are a group of TSOs that are similar to each other but very different from the rest of the sample.
[39] Deuchert, E. and Parthasarathy, S. (2018–19), five-part series of articles on the German energy regulator's benchmarking framework covering efficiency methods (DEA and SFA), functional form assumptions, cost driver analysis, outlier analysis and model validation, *ew–Magazin für die Energiewirtschaft*.
[40] Thanassoulis, E. (1999), 'Setting Achievement Targets for School Children', *Education Economics*, **7**:2, pp. 101–19.

test iteratively until no more super-efficient outliers are identified. This modification is invariant to the volatility of the sample and can result in a more homogeneous sample on which DEA can be performed.

There are further issues with these tests that are not addressed with these amendments. We also note that an outlier procedure is not a replacement for robust data-collection, data-validation and model-development processes.

Sumicsid's second-stage analysis is incapable of detecting omitted cost drivers and does not support the final model.

In order to test whether relevant drivers of expenditure have been omitted from the final model specification, Sumicsid has used regression analysis involving the estimated efficiency scores and the omitted cost drivers.

As noted in Kumbhakar, Parthasarathy and Thanassoulis (2017),[41] we are not aware of any academic literature supporting the use of second-stage regressions to assess the relevance of omitted outputs in a DEA model. In addition, the use of second-stage analysis requires assumptions that need to be justified, and Sumicsid has not presented such justification in its output.

We have demonstrated that Sumicsid's second-stage analysis is unable to validate its own model specification on multiple regulatory datasets. For example, in our review of TCB18 electricity, we demonstrated that Sumicsid's second-stage analysis would be unable to support its own model. For instance, we found that Sumicsid's approach identified only one output as a relevant omitted variable (transformer power) and no other outputs were detected. Regardless of whether similar conclusions would be drawn using the TCB18 gas model, the second-stage analysis is clearly not an appropriate validation procedure and cannot be used to conclude that no relevant cost drivers are omitted from the model.

The results presented in the dynamic efficiency report cast further doubt on the validity of the model and dataset used in the static analysis.

Dynamic efficiency relates to the ability of the most efficient companies in an industry to improve productivity. Sumicsid did not present any relevant analysis regarding dynamic efficiency in the main report, but has subsequently published a separate report on dynamic efficiency. Sumicsid's DEA model indicates that there has been a frontier *regress* of 1.7% p.a.[42] That is, efficient costs have been increasing at a rate of 1.7% p.a. over the period of assessment (i.e. 2013–17).

Such a large and negative frontier shift result is unusual when compared with what is commonly estimated in regulatory settings, and this could indicate that Sumicsid's model cannot capture changes in costs over time—for example, relevant cost drivers that control for the position of a TSO in the investment cycle (such as asset age or quality) are missing.

Sumicsid heavily caveats the dynamic analysis, yet does not do so in its static analysis. This is despite the fact that Sumicsid used the same model and estimation approach and a similar dataset for both the static and the dynamic analysis. For example, Sumicsid states that the frontier shift parameters are

[41] For example, see discussion in Kumbhakar, S., Parthasarathy, S. and Thanassoulis, E. (2018), 'Validity of Bundesnetzagentur's cost driver analysis and second-stage analysis in its efficiency benchmarking approach', February.
[42] Sumicsid (2020), 'Project CEER-TCB18 Dynamic efficiency and productivity changes for gas transmission system operators MAIN REPORT', April.

'affected by a low number of peers in each year (3-4)'.[43] However, it provides no such caveat to its static efficiency analysis, despite also estimating a low number of peers (in the static case there are four peers). It cannot be that the low number of peers limits the robustness of the dynamic analysis but not the static analysis.

[43] Sumicsid (2020), 'Project CEER-TCB18 Dynamic efficiency and productivity changes for gas transmission system operators MAIN REPORT', April, Executive Summary.

1 Introduction

The transmission cost benchmarking project (TCB18), a study of the cost performance of European transmission service operators (TSOs), covering gas[44] and electricity,[45] was commissioned by the Council of European Energy Regulators (CEER) and performed by its consultants, Sumicsid. GTS asked Oxera to perform a critical assessment of the TCB18 gas project.

In gas, TCB18 covered 29 TSOs from 13 European countries, including the 16 German TSOs who were not actively involved in the process and whose data is from a different base year.

Our assessment of TCB18 gas is based on the comprehensive analysis that we have undertaken on behalf of all the electricity TSOs that participated in TCB18,[46] and on a review of all documents published in TCB18 gas. These include the final report and associated appendices, and the slides provided as part of workshops one to five. While we did not have access to the full dataset used by Sumicsid in TCB18 gas, as we had for our assessment of TCB18 electricity, given that the methodologies followed are very similar between the two studies we were able to validate our critique using material in the public domain and the data provided to Sumicsid by GTS.

Through the extensive analysis in this report, we highlight several significant flaws with TCB18 gas. We offer recommendations on the interpretability of the study results and on the additional research required to address the fundamental issues with Sumicsid's analysis.

This report is structured as follows:

- section 2 outlines our assessment of Sumicsid's transparency;
- section 3 critically examines Sumicsid's data-collection and data-construction exercises;
- section 4 assesses Sumicsid's approach to model development;
- section 5 reviews Sumicsid's application and validation of its final model;
- section 6 concludes.

A brief factual summary of TCB18 gas is presented in Appendix A1 and an overview of the recommendations outlined throughout our report is presented in Appendix A2.

[44] Sumicsid (2019), 'Pan-European cost-efficiency benchmark for gas transmission system operators', July.
[45] Sumicsid (2019), 'Pan-European cost-efficiency benchmark for electricity transmission system operators', July.
[46] Oxera (2020), 'A critical assessment of TCB18 electricity', April.

2 Transparency

Transparency was a key issue in our review of TCB18 electricity. Sumicsid's outputs do not contain the necessary information for third parties to clearly follow its analysis, validate its analysis or validate its sources without considerable effort.

We understand that the TCB18 data is confidential and could thus not be published in full. However, public availability of data is only one aspect of transparency; there are several additional aspects that should have been considered to ensure sufficient transparency, like the publication of analysis files or a clear description of processes followed. If those aspects are properly executed, the data would not need to be published to ensure third parties could have a sufficient understanding of the analysis.

Specifically, in our assessment of TCB18 electricity, we observed that the approach chosen and analysis undertaken must be presented in a way that enables easy comprehension and validation. In addition to a clear and comprehensive description of the work, this could take the form of a publication of anonymised analysis files (e.g. a NormGrid[47] Excel calculator) and modelling codes. We have provided an example of such code as part of our review of TCB18 electricity.[48] This is considered practice in regulatory applications, even when the data itself is deemed confidential.[49] For instance, the German regulator, Bundesnetzagentur, published the modelling code for the sectoral productivity factor on its website, even though the data was removed from the website following a court order.[50]

The final report and its appendices lack a clear and comprehensive description of the work and analysis undertaken. Oftentimes the statements Sumicsid makes to support its strong assumptions are not supported by any statistical, operational or literary evidence. Even where Sumicsid states that its assumptions are supported by, for example, statistical evidence, the evidence itself was not shown (e.g. in the form of regression tables).

To illustrate this deficiency we provide an example from TCB18: Sumicsid's 'returns to scale' assumption. Returns to scale (RTS) relates to how changes in inputs (i.e. TOTEX) are linked to changes in outputs (e.g. NormGrid) for efficient companies.

- The returns to scale assumption is one of the key inputs into a DEA model and is mentioned in one slide of the first workshop.[51] No indication as to what RTS assumption is going to be applied in TCB18 is given in this workshop, nor is it discussed how the appropriate RTS assumption will be determined.
- Despite being a key input in the DEA, the returns to scale assumption is then not mentioned again in the material of the subsequent workshops 2–5.

[47] NormGrid and its construction is outlined in appendix A1.
[48] Oxera (2020), 'A critical assessment of TCB18 electricity', April.
[49] Ofwat (2019), 'Final determinations models', available at: https://www.ofwat.gov.uk/final-determinations-models/, NVE (2018), 'IRiR - a script for calculating Revenue Caps for Norwegian Electricity DSOs', available at: https://github.com/NVE/IRiR
[50] Bundesnetzagentur (2018), 'Festlegung des generellen sektoralen Produktivitaetsfaktors fuer Elektrizitaetsversorgungsnetze (3. Regulierungsperiode)', available at: https://www.bundesnetzagentur.de/DE/Service-Funktionen/Beschlusskammern/BK04/BK4_76_Prodfakt/BK4_Prodfakt_Strom_basepage.html
[51] Sumicsid (2018), 'CEER-TCB18 project Methodological Approach', January, slide 16.

No project participant could have been aware of the returns to scale chosen ahead of the publication of the final report.

- The final report then includes the following statement, which is an **exact copy of the text used in e2gas**[52] (the predecessor study to TCB18 gas) and TCB18 electricity.

> For **all possible model specifications**, we have also tested which of the returns to scale assumptions in the DEA model fit data the best […]. We have done so **using F-tests** based on a goodness-of-fit measure as explained in the Method chapter. The general finding is that the IRS assumption […] is the best assumption to invoke. This is **supported also by parametric analyses for a logarithmic model**, where the coefficients sum to less than one for the selected parameters.[53, 54] [emphasis added]

- In e2gas the results of a logarithmic regression were presented alongside this statement.[55] However, this was not the case in TCB18.[56]

From the published material, one would be led to believe that the RTS assumption was made based on strong statistical evidence, despite the fact that the evidence is not presented. It should not be up to the reader to trust Sumicsid's statements—Sumicsid must support its choices with evidence rather than rely on statements alone.

Indeed, in our assessment of TCB18 electricity, we found that we **could not find any statistical evidence supporting the returns to scale assumption** invoked by Sumicsid using the exact same tests described. In contrast we found some evidence supporting a variable returns to scale assumption.[57] Since our analysis was conducted on the same dataset which Sumicsid used it is unclear to us how Sumicsid could have reached a different conclusion.

In response to our report on TCB18 electricity, Professor Per Agrell, on behalf of Sumicsid, stated in the meeting on 5 June 2020 with the ACM that statistical considerations were not the primary reason behind the increasing returns to scale assumption, rather the decision was based on a supposedly established consensus in the literature**.**[58] That is, according to Professor Agrell, the statistical evidence that was stated as underpinning the decision to apply an NRDS (IRS) assumption to the TCB18 data appears to have actually been secondary to theoretical considerations. It is unclear how a project participant, let alone a third party not directly involved in the project (such as a valuations agency), could have discerned this from the report.

The type of transparency outlined above relates to the ability for TSOs and third parties to follow exactly what Sumicsid has done, such that alternatives or improvements to the analysis can be proposed and the limitations of the analysis can be understood. However, in addition to this type of transparency,

[52] Sumicsid and Swiss Economics (2015), 'PROJECT E2GAS Benchmarking European Gas Transmission System Operators – Final report', section 5.32.
[53] Sumicsid (2019), 'Pan-European cost-efficiency benchmark for gas transmission system operators', July, section 5.12.
[54] The tests mentioned are an F-Test based on the estimated efficiencies in the sample (equivalent to the dominance test of the Bundesnetzagentur outlined in section 5.3) and a test of the sum of coefficients in a logarithmic model.
[55] Sumicsid and Swiss Economics (2015), 'PROJECT E2GAS Benchmarking European Gas Transmission System Operators – Final report', Figure 5-3.
[56] Sumicsid (2019), 'Pan-European cost-efficiency benchmark for gas transmission system operators', July, section 5.12
[57] Oxera (2020), 'A critical assessment of TCB18 electricity', April, section 5.1.3.
[58] This is based on our notes as the official transcript of the meeting is not available at the time of writing.

Sumicsid's outputs do not contain key outputs of the benchmarking model that would be fundamentally useful for TSOs and NRAs.

In particular, information on peer companies[59] and importance of different cost drivers to the TSO (virtual weights) are missing entirely. These outputs are essential in helping the TSOs (and the NRAs) in understanding what drives the (in)efficiency of the TSO in the model. For example, a TSO may be interested in understanding whether it is being compared to a TSO with a strong distribution component or a pure point-to-point transport TSO. Only through these outputs can the TSO learn from the benchmarking exercise and thus improve its efficiency by comparing itself to peers. Nevertheless, these outputs are not given to the TSOs in TCB18.

There are many other occasions in which Sumicsid's level of transparency was insufficient, including:

- the treatment of German TSO data (discussed in section 3.1);
- the overarching approach to model development (discussed in section 4.1);
- how the weights on asset classes were derived to estimate NormGrid (discussed in section 4.2);
- the environmental adjustments made (discussed in section 4.3).

We will point these occasions and any other misleading statements made by Sumicsid throughout this report.

[59] Note that this would not necessarily entail naming the efficient TSOs (which could breach confidentiality), but rather describing them qualitatively and quantitatively.

3 Data collection and construction

3.1 Data errors

3.1.1 Description of the issue

Any empirical analysis relies on the accuracy of the data being used, and real data is always subject to some level data errors. These errors can, for example, result from issues such as misreporting, miscommunication and measurement errors. These are empirical issues and their effect on the dataset needs to be examined.

While we do not have access to the full TCB18 gas dataset, we confirmed that all of the above scenarios occurred in TCB18 electricity and were present in the final dataset on which the published results were estimated. Given that TCB18 gas involved the same data-validation procedure and many of the parties responsible for data validation (e.g. NRAs, Sumicsid) are the same, similar errors are also likely to exist in the TCB18 gas data.

TCB18 gas faces an additional data-quality issue (when compared with TCB18 electricity) due to the inclusion of the German gas TSOs' data, which makes up the majority of the sample in the study. The German gas TSOs did not participate in the TCB18 process beyond their attendance of the initial workshop, and it is our understanding that they therefore had minimal opportunity to ensure the accuracy of their data. Their data was from an earlier period to that considered for the other participating TSOs, and adjustments were needed to make it comparable to the data collected based on the TCB18 guidelines.[60] Both of these issues add a further level of uncertainty to the German TSOs' data.

While some methods are better able to account for data inaccuracies than others, DEA as applied by Sumicsid in TCB18 does not account for them at all. Sumicsid's DEA application is particularly sensitive to data errors as it results in a deterministic method of estimation failing to account for statistical noise (such as random data errors and modelling uncertainty). In other words, the full gap between a TSO and the estimated frontier is considered as inefficiency. Efficiency estimates also exclusively rely on the identified benchmark (i.e. efficient) TSOs. Therefore, any TSO identified as a benchmark and placed on the efficient frontier because of a data error will directly bias the estimated efficiencies for other TSOs for which this particular TSO acts as a benchmark. The reverse is also true, in that a TSO failing to be identified as a benchmark through a data error could also affect the position and shape of the efficient frontier.[61]

Therefore, a rigorous data-screening process is required before any empirical assessment including model development or efficiency analysis is performed.

[60] Sumicsid (2019), 'Pan-European cost-efficiency benchmark for gas transmission system operators main report', July, section 4.101. Note that, as the data for the German TSOs was reported in 2015 while the rest of the sample's data was reported in 2017, there need to be additional adjustments to correct for this temporal inconsistency.
[61] Placing a TSO by error on the efficient frontier would adversely affect the efficiencies of TSOs for which it is a peer. Similarly, failing to place a TSO by error on the efficient frontier would benefit the efficiencies of TSOs for which it might have been a peer. As data errors can be random, in a comparative assessment we cannot know which TSOs benefited and which suffered in their comparative efficiency through TSOs being erroneously placed on the frontier or, conversely, through failing to feature on the frontier.

3.1.2 Sumicsid's approach

As outlined in detail in Appendix A1, Sumicsid states that its data collection and validation procedure consisted of six stages.[62] Sumicsid states that 'although no approach will be fully safe' the datasets are of 'good quality'.[63]

Regarding the German TSO data, Sumicsid states that it is satisfied that 'particular attention was paid to screen and analyze any potential differences between the reporting instructions in the previous benchmark, the national validation performed and potential sources of errors'.[64] It concludes that 'This latter process resulted in a positive analysis for the gas TSO, deemed to offer comparable data of high validated quality'.[65]

3.1.3 Critique and proposed solution

Issues in TCB18 data validation

Sumicsid followed an identical data collection and validation[66] process in gas as it did in electricity, and its description of the overall procedure to data collection suggests that it has followed a process that appears relatively consistent with good practice. Despite this, multiple TSOs in the electricity shadow study noted that their data used by Sumicsid was not accurate. Some electricity TSOs flagged data inaccuracies to Sumicsid over the course of the study, but these were not corrected in its final analysis. In other cases, data was clearly erroneous and in violation of Sumicsid's stated validation criteria. For example, Sumicsid's criteria for data validation in electricity stated that the proportion of angular towers should be at least 10%,[67] yet one TSO in the electricity study did not report any angular towers. This is indicative of wider issues in the data collection and suggests that the six-stage data collection and validation approach may not have been followed rigorously.

Even if the stated process was followed, national data validation by the NRAs may not be sufficient to ensure a consistent international dataset. It is not clear if and to what degree the NRAs followed common guidelines or proceeded with the same level of care in scrutinising the data. If each NRA only had access to its TSO's data and followed its own assessment framework, it is likely that the resulting sample would be inconsistent and contain more errors than if knowledge of data issues was shared and openly discussed with the participants.

For example, we understand that the ACM excluded all of GTS's connection points with a flow rate below a certain cut-off from the asset dataset. However, it is not stated in the reporting guidelines or the main report that this approach was to be taken by all participating NRAs.[68] Regardless of whether these assets *should* be included in the dataset, if the approach taken to data validation differs by NRA, then the resulting dataset will be inconsistent and not suitable for benchmarking. In this instance, if some regulators did not remove

[62] Sumicsid (2019), 'Pan-European cost-efficiency benchmark for gas transmission system operators main report', July, section 3.2.
[63] Sumicsid (2019), 'Pan-European cost-efficiency benchmark for gas transmission system operators main report', July, sections 3.06–3.07.
[64] Sumicsid (2019), 'Pan-European cost-efficiency benchmark for gas transmission system operators MAIN REPORT', July, p. 7.
[65] Sumicsid (2019), 'Pan-European cost-efficiency benchmark for gas transmission system operators MAIN REPORT', July, p. 7.
[66] With the exception of additional data collection and validation procedures for the German TSOs.
[67] This was stated by Sumicsid in a validation document for one of the electricity TSOs. See Sumicsid (2018), 'TCB18 Crossvalidation TECH/ELEC', October, p. 1.
[68] CEER and Sumicsid (2018), 'CEER TSO Cost Efficiency Benchmark - Gas asset reporting guide', March, p. 10.

connection points from the asset data, the results of the study will be biased against those TSOs for which the NRA removed assets from the benchmark, such as GTS, and biased in favour of those that did not.

The issue of inconsistent interpretation of reporting guidelines is not limited to NRAs. In the final report, Sumicsid argues that few TSOs reported significant rehabilitations[69] and, of the TSOs that did, some reported this incorrectly.[70] Sumicsid chose to exclude significant rehabilitations from the benchmark 'in order not to compromise data quality'.[71] This is an additional example of where project participants have not consistently interpreted the data-reporting guidelines, and this introduces another layer of uncertainty in the data that could bias the analysis against some TSOs. Indeed, the failure to account for significant rehabilitations could bias the analysis against TSOs with an old network (such as GTS), as these TSOs are more likely to have undertaken significant investments to extend the lives of old assets.

Issues arising from the inclusion of German TSO data

Unlike in the case of TCB18 electricity, for TCB18 gas, the German TSOs (the majority of the TSOs in the sample) did not participate in the TCB18 process beyond the initial workshop. Sumicsid states that 'particular attention was paid to screen and analyse any potential differences between the reporting instructions in the previous benchmark, the national validation performed and potential sources of errors'.[72] However, Sumicsid provides no description of any screening process for data issues and differences in reporting which took place. This is another instance of Sumicsid's transparency falling short of what would be considered good practice.

While the precise process in place is not described, it is our understanding from communication with GTS that the German TSOs were not involved in the data validation. In validating data and spotting inconsistencies, the input of the TSOs is invaluable as they have a better understanding of operational expectations of asset characteristics. Thus, regardless of any process that took place, the possibility of data errors is higher without the oversight of or input from the TSOs.

There is also no record of any differences in reporting or assumptions between TCB18 data and the German gas TSOs data or how these differences were corrected for. For instance, German regulatory benchmarking of gas TSOs does not consider the amalgamation of assets into a 'NormGrid' type variable as a cost driver. Thus, the availability and granularity of asset-level data of the German TSOs remain unexplained in the study.

In fact this is not the case. Sumicsid explicitly states that 'compressor power per individual compressor' was not available for German gas TSOs,[73] yet this

[69] Significant rehabilitations refers to investments in existing assets designed to prolong their useful lifetimes, rather than to upgrade their capacity. As the measured level of output does not change with this increase in investment, it may increase a company's inefficiency regardless of how efficient the investment is. For this reason, adjustments need to be made to the investment data if TSOs are to be benchmarked in an unbiased manner.
[70] Sumicsid (2019), 'Pan-European cost-efficiency benchmark for gas transmission system operators main report', July, p. 37.
[71] Sumicsid (2019), 'Pan-European cost-efficiency benchmark for gas transmission system operators main report', July, p. 37.
[72] Sumicsid (2019), 'Pan-European cost-efficiency benchmark for gas transmission system operators MAIN REPORT', July, p. 7.
[73] Sumicsid (2019), 'Pan-European cost-efficiency benchmark for gas transmission system operators main report', July, p. 32.

is an input in the calculation of NormGrid.[74] Therefore, it appears that Sumicsid made at least some assumptions when calculating its cost drivers from the German TSO data. Any assumption or estimation made for this will invariably introduce some uncertainty into the data and may bias results.[75]

In addition to the comparability of the data (in terms of reporting consistency), the German data is reported for the year 2015, whereas the data for other TSOs is reported in 2017. As Sumicsid's dynamic efficiency analysis indicates that efficient costs in the industry are rising at a rate of 1.7% p.a., this temporal discrepancy could bias the analysis in favour of the German TSOs, at the expense of others.

Possible adjustments for data issues in regulatory benchmarking

Regulators regularly make adjustments to reflect modelling limitations, data errors and other informational deficiencies. For example, in the UK, the water and energy regulatory authorities (Ofwat and Ofgem, respectively) have focused on the upper quartile rather than the efficiency frontier in order to account for data and modelling uncertainties.[76]

Regulators have also considered triangulating their assessment against the firm's view in recognition of informational limitations, thereby reducing the challenge further.[77] Similarly, the UK rail regulator (the Office of Rail and Road, ORR) has made a downward adjustment to its estimated efficiencies of 25%, having used a regression approach[78] (which, like DEA, assumes that the entire gap to the estimated cost frontier is due to inefficiency). Other regulators have used similar adjustments—for example, by benchmarking to the upper-decile position even accounting for statistical noise explicitly using SFA.[79] The Bundesnetzagentur uses the 'best of' results of DEA and SFA to give the benefit of the doubt regarding companies' efficiency levels, with the latter method accounting for such uncertainty explicitly in the modelling.[80]

Thus, even when regulators in a national benchmarking context follow good scientific practice in terms of data collection, model development and model validation, they typically apply downward adjustment to account for general errors and uncertainties in top-down cost benchmarking. The magnitude of the adjustment for noise should take into account the available evidence from approaches such as Monte Carlo simulation and SFA.

[74] Sumicsid (2019), 'Pan-European cost-efficiency benchmark for gas transmission system operators main report – NormGrid development', p. 17.
[75] Further explicit differences in reporting are clearly present. For instance, in German regulatory benchmarking of gas TSOs no correction for ownership is made for connection points or pipe length. However, according to the TCB18 asset reporting guidelines the usage share of any asset must be reported. Swiss Economics, Sumicsid and 4-Management (2017), 'Kostentreiberanalyse und Effizienzvergleich der Gasfernleitungsnetzbetreiber Arbeitsstand Parameterdefinitionen', p. 6; and CEER (2018), 'CEER TSO Cost Efficiency Benchmark - Gas asset reporting guide', p. 5.
[76] For example, see Ofgem (2013), 'RIIO-ED1 business plan expenditure assessment - methodology and results', December, p. 8. Even the use of an upper-quartile benchmark was challenged as being overly stringent in the courts. For example, in its re-determination of Bristol Water's cost allowance at PR14, the UK Competition and Markets Authority revised the benchmark down to the average. See Competition and Markets Authority (2015), 'Bristol Water plc: A reference under section 12(3)(a) of the Water Industry Act 1991', October.
[77] For example, the Norwegian revenue cap model for network companies (https://www.nve.no/norwegian-energy-regulatory-authority/economic-regulation/). See also Ofgem (2014), 'RIIO-ED1: Final determinations for the slowtrack electricity distribution companies', November, p. 8.
[78] Office of Rail and Road (2018), 'PR18 Econometric top-down benchmarking of Network Rail', July, pp. 8–9.
[79] For example, see NERA (2008), 'The Comparative Efficiency of BT Openreach: A report for Ofcom', March, p. 51.
[80] This is explicitly outlined in the law (see ARegV §13).

In our review of TCB18 electricity, we used a Monte Carlo simulation to estimate the impact of data errors.[81] The Monte Carlo simulation adds a random component to all inputs and outputs,[82] and the magnitude of this random component was informed by our knowledge on the prevalence of data errors in the TCB18 dataset. Using the Monte Carlo simulation, we found that error margins of 10–18% might be appropriate in electricity, based on data uncertainty alone.[83] Moreover, when applying SFA to Sumicsid's electricity specification, we found that no statistically significant inefficiency existed on the model.[84] In other words, all of a TSO's gap to the cost frontier that was estimated from the model could be explained by modelling uncertainty and random data errors.

Without access to the final data, it is not possible for us to model the impact of data errors on GTS's estimated efficiency score. However, given the significant additional uncertainty in the TCB18 gas data introduced by the inclusion of German gas TSO data, a similar outcome could be possible. That is, much (or all) of the estimated efficiency gap in TCB18 gas could be explained by statistical noise.

3.2 Defining the input variable

3.2.1 Description of the issue

Sumicsid uses TOTEX as the single input in its model—its measure of TOTEX in a given year is constructed as the sum of OPEX in a given year and the sum of CAPEX annuities leading up to that year. For this approach to provide the correct incentives and a robust estimate of managerial efficiency, it must be assumed that the relationship between benchmarked OPEX and benchmarked CAPEX are as follows.

- Equivalent—at the margin, €1 of CAPEX should have the same worth as €1 of OPEX in terms of supporting output levels. That is, there must be a one-to-one trade-off between OPEX and CAPEX.
- Controllable—OPEX and CAPEX as well as the ratio between the two need to be within the control of management.

If the above conditions do not hold, then a TOTEX model will be inappropriate and could lead to unrealistic efficiency challenges for some TSOs. For example, if a TSO does not own the assets it operates (leasing them from another entity instead), it will have significantly lower level of CAPEX compared with TSOs that do own the assets. Using a TOTEX model, TSOs that do own the assets may be benchmarked against the TSO that does not, and the former group of TSOs may have a large estimated inefficiency because of this. However, the TSOs that do own their assets cannot replicate the performance of the TSO that does not, meaning that the estimated cost savings are not feasible. This is not simply a theoretical argument—it occurred in TCB18 electricity.[85]

[81] For example, an error discovered in a TSO's CAPEX overstates its TOTEX by 32%. In another case, the value for weighted lines is overstated by 27%. See Oxera (2020), 'A critical assessment of TCB18 electricity', Prepared on behalf of European electricity TSOs, April, section 3.1.
[82] For example, if a TSO reported a TOTEX of €10m then we would draw a number between €9m and €11m from a uniform distribution. We do the same thing for all other outputs and for all TSOs. We then estimate the efficiencies. We repeat this process 1,000 times to arrive at a distribution of estimated efficiency scores.
[83] That is, even if the model was correctly specified and robustly validated, the model cannot predict TSOs' efficiency score with a high degree of certainty. It is likely that a larger adjustment than 18% may be required if there are aspects of the modelling that systematically bias the model against specific TSOs.
[84] Oxera (2020), 'A critical review of TCB18 electricity', April, section 5.5.3.
[85] Oxera (2020), 'A critical assessment of TCB18 electricity', April, p. 37.

3.2.2 Sumicsid's approach

Sumicsid uses a TOTEX model to assess the efficiency of TSOs. It states that such a model provides incentives for TSOs to balance OPEX and CAPEX solutions,[86] but the equivalence and controllability conditions are not discussed in its report.

3.2.3 Critique and proposed solution

Sumicsid has made separate adjustments and normalisations to OPEX and CAPEX, as shown in Table 3.1.

Table 3.1 Sumicsid's cost normalisation approach

Issue	OPEX	CAPEX
Time period over which it is assessed	Annual (i.e. only 2015 OPEX is considered for German TSOs, and 2017 OPEX for the rest of the sample)	A sum of annuitized investments from 1973 and an opening investment estimate
PLI adjustment	Adjustment for labour costs using the PLI for civil engineering works	No PLI adjustment
Inflation adjustment	German TSO data adjusted for inflation in overall goods [1]	Adjusted for inflation in overall goods

Note: [1] Sumicsid's final model is estimated using OPEX from the year 2017 only for all other TSOs. As such, the inflation adjustment is not applicable.

Source: Sumicsid (2019), 'Pan-European cost-efficiency benchmark for gas transmission system operators main report'.

Operationally, it is unlikely that €1 of CAPEX should have the same impact as €1 of OPEX—at the margin, €1 of OPEX will not have the same impact on output levels (i.e. in Sumicsid's model, assets) than €1 of CAPEX. Moreover, the separate treatment of OPEX and CAPEX in the cost normalisation process casts doubt on the equivalence between the two. A manager could re-allocate €1 from *actual* OPEX to *actual* CAPEX and the resulting *normalised* TOTEX will be different. Similarly, two TSOs that have equivalent levels of TOTEX may have different level of *normalised* TOTEX due to differences in cost reporting.

Due to the normalisation of CAPEX outlined in Table 3.1, benchmarked CAPEX in TCB18 is different from the actual CAPEX on investments in a year incurred by the TSO. In fact, the majority of CAPEX in TCB18 for GTS was incurred before 1992. Therefore, it is clear that CAPEX is not fully within the control of the current management team at GTS. As CAPEX is not fully controllable, TSOs cannot simply re-allocate a unit of CAPEX to OPEX. Therefore, any inefficiencies that may exist due to allocation issues between the cost categories (e.g. if the TSO previously spent too much CAPEX as opposed to OPEX) cannot be reduced and should not be part of a cost reduction target.

In this context, it is inappropriate to impose a rigid one-to-one trade-off between OPEX and CAPEX by treating TOTEX as a single input. Instead, to ensure that an efficiency estimate is appropriate as a cost reduction target, one option would be to consider OPEX and CAPEX as separate inputs in the DEA model.[87] This would directly account for the heterogeneity between the cost

[86] Sumicsid (2019), 'Pan-European cost-efficiency benchmark for gas transmission system operators main report', July, p. 19.
[87] Multi-input models can also be estimated in an SFA context. See Kumbhakar, S.C, Wang, H-J. and Horncastle, A.P. (2015), *A Practitioner's Guide to Stochastic Frontier Analysis Using STATA*, Cambridge University Press, chapter 6.

categories and allow a cost reduction target to be set on only the controllable component.

Such an approach would still assume that TSOs are freely able to reduce their measured CAPEX. However, CAPEX (as defined by Sumicsid) represents the historical investments of the TSO and is therefore the ‘sunk costs’ that TSOs cannot readily reduce. Therefore, an alternative approach to a simple two-input model would be to treat CAPEX as an exogenously fixed input. This way, the estimated efficiency will simply reflect the potential reduction in the controllable inputs (e.g. OPEX).

In the meeting with Sumicsid and the ACM, Sumicsid stated that developing a two-input model would mean that TSOs’ estimated efficiency scores would be driven by arbitrary weights on OPEX and CAPEX derived from the DEA model. However, Sumicsid did not acknowledge that its approach already imposes arbitrary weights on OPEX and CAPEX. Furthermore, if Sumicsid is concerned that the weights estimated in the DEA model could imply unrealistic trade-offs between OPEX and CAPEX, it is possible to introduce weight restrictions. This approach ensures a more flexible relationship between the cost categories. Note that Sumicsid has not assessed the estimated weights on the *output* variables in its model with the same level of scrutiny as it has the estimated weights on *input* variables.

An alternative approach to a two-input model could be to construct TOTEX as a weighted average of OPEX and CAPEX (where the weights are based on expert judgement), recognising that some TSOs are able to make trade-offs, but maintaining flexibility regarding the exact relationship.[88]

Furthermore, it is also possible to model activities at a disaggregate level (e.g. developing separate OPEX and CAPEX models or separate models for transport, maintenance and planning) before aggregating to a TOTEX efficiency. This is one of the approaches used by UK regulators.[89] Such models could also serve as a cross-check for models developed on a TOTEX basis as they can capture drivers of specific types of costs more robustly.

The exact impact on GTS’s efficiency depends on the approach taken and the effects on the outlier analysis. However, we note that in TCB18 electricity, moving to an (unrestricted), two-input DEA model improved the estimated efficiency of some TSOs by up to 17%, and it is therefore a highly material issue. Moving to a two-input model also meant that electricity TSOs were benchmarked to more comparable peer companies (in terms of operating structure). This may be a material issue for GTS, as its efficiency score is highly sensitive to changes in the WACC, indicating that it may be more CAPEX-intensive than its peers. Indeed, reducing the WACC parameter from 3% to 1.8% increases GTS’s estimated efficiency from 73.2% to 81.4%. We understand that the ACM is intending to apply a 0.7% WACC parameter when setting allowed revenues for GTS in the next regulatory period. The ex ante expectation is that reducing the WACC parameter to this level would further

[88] The weighted average representation of TOTEX would be: TOTEX = w*CAPEX +(1-w)*OPEX = OPEX +w*(CAPEX-OPEX). The weight w can be estimated through a regression of OPEX on CAPEX. For additional flexibility, a squared or cubed term can be included. Typically, there is a confidence interval around the weight and it can be introduced in the two-input DEA model in the form of an additional constraint specifying the relationship between the two.
[89] Ofwat (2019), ‘PR19 final determinations: Securing cost efficiency technical appendix’, December; and Ofgem (2019), ‘Consultation - RIIO-2 tools for cost assessment’, June.

increase GTS's estimated efficiency beyond 81.4% under Sumicsid's framework.[90]

3.3 Price levels

3.3.1 Description

Price-level differences persist even in closely linked economies and for relatively mobile goods.[91] It is not possible for a TSO to fully mitigate the impact of higher input prices. That is, input prices are an exogenous factor (i.e. not within management control).

Controlling for the impact of input prices on expenditure is an important step in normalising costs over different jurisdictions, and is often done in national and international benchmarking studies.[92]

3.3.2 Sumicsid's approach

Sumicsid acknowledges that price-level differences exist for some input factors, namely labour. It further concedes that input prices are exogenous:

> In order to make the operating costs comparable between countries a correction for differences in national salary cost levels has been applied. Otherwise TSOs would be held responsible for cost effects, e.g. high wage level, which is not controllable by them.[93]

In particular, Sumicsid applies a correction to direct manpower costs within OPEX using the PLI for civil engineering from Eurostat (referred to as 'Plici' by Sumicsid). The index 'includes construction not classified under buildings, for example railways and bridges'.[94]

3.3.3 Critique and proposed solution

Sumicsid correctly recognises the need to correct for price-level differences across TSOs. However, we have identified two broad issues with Sumicsid's approach.

- Sumicsid does not justify using the PLI for civil engineering for direct manpower costs with statistical or operational evidence. Using a PLI for civil engineering does not take into account the fact that labour costs, as an immobile production factor, vary by more than overall civil engineering price levels, which includes factors of production that are more mobile across

[90] Note that we do not consider the adjustment to the WACC parameter to be a reasonable solution to all of the issues with TCB18 that we have identified in this report. That is, reducing the WACC to the level used by the ACM is still not a robust basis on which to set regulated revenues.
[91] For studies of price levels in closely linked economies, see, for example, Berka, M. and Devereux, M.B. (2010), 'What determines European real exchange rates?', National Bureau of Economic Research; and Engel, C. and Rogers, J.H. (1996), 'How wide is the border?', *The American Economic Review*, **86**:5, pp. 1112–25. For studies on the effect of a single currency, see, for example, Engel, C. and Rogers, J.H. (2004), 'European product market integration after the euro', *Economic Policy*, **19**:39, pp. 348–84; and Eurostat (2019), 'GDP per capita, consumption per capita and price-level indices', https://ec.europa.eu/Eurostat/statistics-explained/index.php?title=GDP_per_capita,_consumption_per_capita_and_price_level_indices#Relative_volumes_of_GDP_per_capita, accessed 26 November 2019.
[92] For example, see Office of Rail and Road (2013), 'PR13 Efficiency Benchmarking of Network Rail using LICB', August, pp. 12–14.
[93] Sumicsid (2019), Pan-European cost-efficiency benchmark for gas transmission system operators main report, July, p. 20, section 4.50.
[94] Eurostat (2018), Glossary: Civil engineering work, available at: https://ec.europa.eu/Eurostat/statistics-explained/index.php/Glossary:Civil_engineering_work, accessed 8 December 2019; and Eurostat (2019), Purchasing power parities (PPPs), price-level indices and real expenditures for ESA 2010 aggregates, available at: https://appsso.Eurostat.ec.europa.eu/nui/show.do?dataset=prc_ppp_ind&lang=en, accessed 29 November 2019.

borders such as raw materials (e.g. construction materials such as metals, plastics and concrete).

- Sumicsid does not apply a correction to any cost item other than direct manpower cost in OPEX. This means that Sumicsid only accounts for input price differentials for a very small proportion of the cost base (for GTS direct manpower cost constitutes just 6% of TOTEX) without sufficient justification. As such, Sumicsid assumes that there are no input price differences across the participating TSOs for all the other components of OPEX, such as purchase of external maintenance, personnel leasing, consultancies, office supplies and control centre costs, as well as all components of CAPEX including labour costs.

The adjustment made by Sumicsid is therefore insufficient as it does not capture all of the material differences in price levels between countries. In fact, the TCB18 study assumes that maintenance services and all investment goods can be procured for the same price in the Netherlands as they can be in Finland, Slovenia and the UK.[95] Sumicsid does not provide evidence to validate its hypotheses, nor does it consider that:

- a significant proportion of CAPEX is labour or labour-related costs (e.g. installation costs) and not covered by the adjustment to gross labour costs in OPEX. In e2gas for example 30% of CAPEX was considered labour cost and adjusted for price levels using the EUROSTAT EU salary index.[96] Sample projects provided by GTS estimate the share of labour in project costs at 54–63%, and Sumicsid itself assumes a 41.5% share of 'installation costs' in total costs of pipelines; [97]
- differences in the price of more mobile factors of production, such as raw materials, do exist across the EEA due to, for example, transportation costs, as evident in differences in the investment goods, total goods and other price indices;[98]
- much of the (44-year) CAPEX investment stream used to calculate TOTEX was incurred decades ago, before the close integration of the EU and EEA was complete. For instance, GTS incurred nearly 50% of its CAPEX in TCB18 prior to 1990 and thus well before the free movement of labour and capital within the EU was introduced. It should be clear that GTS was not part of the same market as the TSOs operating in the Soviet sphere of influence at the time due to stringent trade restrictions.[99] It is therefore unlikely that they faced the same input prices.

The approach chosen by Sumicsid therefore does not sufficiently normalise the cost base, and risks conflating the uncontrollable price-level differences faced by companies with managerial inefficiency.

[95] Note that Sumicsid uses country-specific measures of inflation to adjust historical expenditure. However, this adjusts for differences in prices through time *within* a country, but does not in any way adjust for differences in price levels *across* countries.
[96] Sumicsid and Swiss Economics (2016), 'PROJECT E2GAS Benchmarking European Gas Transmission System Operators - FINAL REPORT', June, p. 23.
[97] Sumicsid (2019), 'Pan-European cost-efficiency benchmark for gas transmission system operators APPENDIX, July.
[98] Eurostat (2019), Purchasing power parities (PPPs), price-level indices and real expenditures for ESA 2010 aggregates, available at: https://appsso.Eurostat.ec.europa.eu/nui/show.do?dataset=prc_ppp_ind&lang=en, accessed 29 November 2019.
[99] Trade within the Soviet sphere of influence was conducted under a different system of exchange. See Broadman, H. G. (ed.). (2006), *From disintegration to reintegration: Eastern Europe and the former Soviet Union in international trade*, The World Bank, p. 52, Box 1.1.

In national and international benchmarking exercises, it is common to adjust all expenditure to account for regional and international differences in prices. In the UK, for example, the Office of Rail and Road (ORR) adjusted all cost data (100% of OPEX and CAPEX) using the PLI for GDP adjustment in its international benchmarking of Network Rail's efficiency for its 'PR13' price review.[100] Differences in input prices are sometimes considered in national benchmarking exercises.[101] This illustrates that price levels are an important issue in any benchmarking exercise, but are particularly important in international comparisons.

Although it is clear that input price adjustments are required across a material proportion of the cost base, if not the entire cost base, the precise method of correcting for price differences requires careful consideration. In particular, one needs to consider:

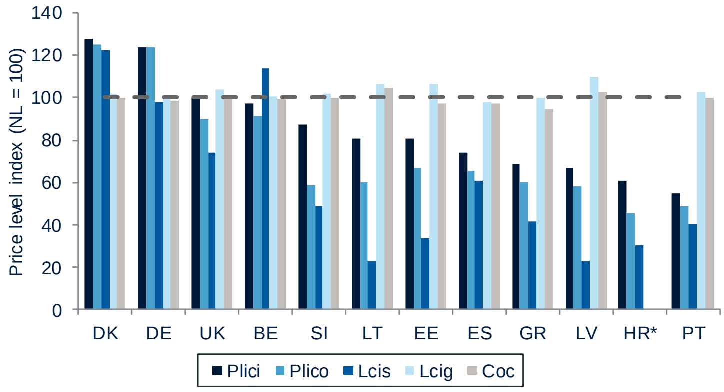
- the basket of goods represented by the PLI (e.g. consumer prices, construction prices);
- the base year of the PLI (e.g. as the published values of price levels can vary across years, should we express prices in price levels for 2017, 2016 or an average over the modelled period?);
- the percentage and type of the cost base subject to the PLI adjustment (e.g. are there cost components, such as within CAPEX, whose prices do not vary across countries or TSOs?);

The estimated efficiency scores for individual TSOs are likely to be highly sensitive to these choices, illustrating the need for sensitivities around the choices.

For example, German TSOs face prices that are 20% higher than GTS according to Sumicsid's selected PLI (Plici), but face broadly similar prices according to three other PLIs (Lcis, Lcig and Coc).[102] All of these price indices were considered by Sumicsid, yet it has not provided any valid justification for choosing Plici.

[100] Office for Rail and Road (2013), 'PR13 Efficiency Benchmarking of Network Rail using LICB', August, pp. 12–14, August; (2008), 'Periodic review 2008 Determination of Network Rail's outputs and funding for 2009-14', October, p. 122. For the latest benchmarking exercise, which is part of PR18, the ORR did not perform an international benchmarking exercise.
[101] For example, in its RIIO-ED1 price control, the Office of Gas and Electricity Markets (Ofgem) applied a correction for regional labour costs within the UK to its cost base. See Ofgem (2014), 'RIIO-ED1 final determinations for the slow-track electricity distribution companies Business plan expenditure assessment', 28 November, p. 41.
[102] Lcis and Lcig are LCIs for services and goods, respectively, while Coc is a PLI for construction.

Figure 3.1 Price level indices (NL = 100)



Note: *Lcig and Coc data are unavailable for Croatia.

Source: GTS analysis of EUROSTAT data.

The method of accounting for price-level differences is clearly not a trivial decision and must be robustly justified. In our review of TCB18 electricity, we noted that one option would be to adjust all OPEX with the PLI for overall GDP and to adjust CAPEX with the PLI for civil engineering, as the activities contained in these indices appear aligned with TSOs activities.

We maintain that, on OPEX, which contains a wide array of goods and services bought by TSOs (e.g. grid maintenance, grid planning and business support activities), the PLI for overall GDP may be a reasonable compromise between the higher price-level differences in services and the lower price-level differences in goods, although further analysis is needed before this index can be applied to TSOs' OPEX.

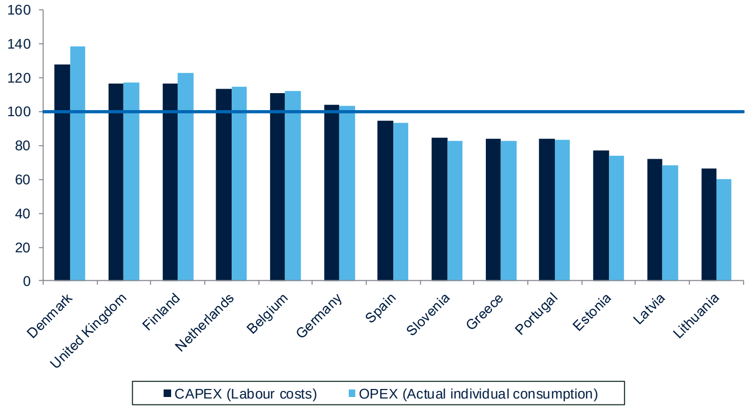
In our review of TCB18 gas, we have considered an alternative adjustment to CAPEX based on labour costs only. For this adjustment, GTS provided us with a breakdown of CAPEX for a typical investment project into activities (e.g. engineering, construction, management) and the associated share of labour related costs these activities contain.[103] Based on this data, we calculated possible adjustments to CAPEX accounting for labour as an immobile factor in CAPEX using the PLI for services, assuming that GTS is a 'typical' TSO in terms of the composition of its cost base. As can be seen in Figure 3.2, this suggests that the price of investment projects can vary by up to 48% depending on price levels alone. A benchmark calculated on unadjusted cost data would thus materially understate (overstate) the efficiency of investments by TSOs operating in a country with a high (low) price levels.

Although GTS is not uniquely affected by this issue, the Netherlands is a relatively high price-level country. Therefore, it is likely that at least part of the

[103] As such a detailed breakdown was not available to us in electricity, we suggested the application of a PLI for civil engineering works instead. Oxera (2020), 'A critical assessment of TCB18 electricity', April, section 3.3.3.

estimated efficiency gap for GTS is a result of uncontrollable relative price levels.[104]

Figure 3.2 Price level differences across the sample



Source: Oxera analysis.

In addition, some sensitivity analysis should be undertaken with respect to the choice of price level adjustment to demonstrate the robustness (or limitations) of the analysis.

As Sumicsid has not sufficiently motivated its choices in its final outputs, it has potentially conflated the estimated efficiency scores with price-level differences.

3.4 Indirect cost allocation

3.4.1 Description of the issue

Like most large businesses, TSOs have overarching support functions (described as 'Indirect Support' by Sumicsid), such as finance, IT support and human resources. Some of these support functions may be directly relevant to specific activities, but others may be sufficiently general that they cannot be allocated to any one activity.

As many TSOs perform activities that are beyond the scope of the TCB18 benchmarking project, some of the costs incurred in 'Indirect Support' will be driven by activities that are not assessed. For example, a TSO that undertakes a significant amount of offshore transport or operates LNG terminals (activities that are outside the scope of benchmarking) may have a larger expenditure on IT support than other TSOs that do not undertake such activities.

Thus, to avoid a TSO's efficiency score being driven by activities deemed outside the scope of the study, indirect costs need to be allocated to activities within the scope of benchmarking. Ideally, the allocation rule should consider how much of the indirect costs are driven by each activity (i.e. where the indirect expenditure is incurred).

[104] The direction of the bias mainly depends on the relative price level between GTS and its peers (which are not known). However, as two of the highest cost TSOs (Gasum and Energinet) are not peers, we consider a downward bias is most likely.

3.4.2 Sumicsid's approach to allocating indirect expenditure

Sumicsid's allocation rule uses the proportion of all costs except depreciation and energy relative to these costs across activities.[105]

$$Indirect_{TPM} = \frac{OPEX_{TPM} - Depreciation_{TPM} - Energy_{TPM}}{OPEX_{total} - Depreciation_{total} - Energy_{total}} * Indirect_{total}$$

where the subscripts TPM and $total$ refer to expenditure in in-scope activities (i.e. transport, planning and maintenance) and all OPEX, respectively. Note that this allocation rule *includes* cost items such as taxes and research and development, which are not part of the efficiency benchmarking.

Sumicsid states that it has tested 'several allocation methods' for indirect expenditure.[106] However, sensitivities regarding the allocation rule were not presented in any of its final reports.

3.4.3 Critique and proposed solution

Sumicsid's allocation rule is generic, not specific to TSOs support expenditure, and never justified in its main report. In some cases, the allocation of indirect expenditure may be driven by large cost items that are unrelated to where indirect costs are incurred. In TCB18 electricity for instance, one TSO paid a large levy that was assigned to an out-of-scope activity. Owing to this, nearly all indirect expenditure for this TSO was allocated to this out-of-scope activity, even though the tax payment itself should not be a material driver of indirect expenditure.

As a solution to this issue, we consider that it would be more robust to use only in-scope cost categories to allocate indirect OPEX. This would mitigate the risk that indirect costs are allocated to activities based on cost items that are unrelated to indirect expenditure. In the electricity study, correcting for this issue reduced one TSO's estimated efficiency from 96% to 79%. If this TSO were estimated to be fully efficient, this issue would have had a significant impact on the estimated efficiency of more TSOs in the sample.

The allocation of indirect costs to in-scope activities varies between 20% and 100% across the gas TSOs. If a peer TSO had incorrectly allocated a large proportion of indirect costs to out-of-scope activities, the impact on other TSOs' estimated efficiency could be substantial.

[105] Sumicsid (2019), 'Pan-European cost-efficiency benchmark for gas transmission system operators main report', July, section 4.103.
[106] Sumicsid (2019), 'Pan-European cost-efficiency benchmark for gas transmission system operators main report', July, section 4.97.

4 Model development

4.1 Cost driver analysis

4.1.1 Description of the issue

Operating, maintaining and enhancing a gas transmission network is an extremely complex operation. Finding a set of cost drivers that can completely describe the functions of a TSO is therefore a difficult task. A robust clearly defined model-development process is essential in ensuring that the results from the empirical analysis are robust and no relevant cost drivers were omitted. This model-development process should take into account both the operational and economic rationale for including specific cost drivers, as well as their statistical validity. In performing the analysis, the assumptions of any statistical model should be justified and, wherever possible, empirically tested.

4.1.2 Sumicsid's approach

In the workshop slides preceding the final report, Sumicsid states that it used three estimation approaches to examine the relationship between costs and cost drivers:

- OLS regression;
- OLS regression excluding outliers as defined by the Cook's distance metric;
- ROLS regression (an estimator where observations far from the regression line are given less weight).

Sumicsid states that it has also used Lasso regression to justify the size of its model (i.e. the number of output variables) and its use of NormGrid. Based on this analysis, Sumicsid concludes that the optimal model size is 2–4 parameters (cost drivers).[107]

The cost driver analysis process is not presented or discussed in the final report. The regression results of Sumicsid's final model is shown in Table 4.1.

Table 4.1 TCB18 final model

	TCB18 model
NormGrid (slope)	1.841***
Compressor Power	−178.510***
Connection Points	207935.700***
Pipe Length (humidity)	−14,286.860***
Constant1	-
R-squared	0.997

Note: [1] Sumicsid suppresses the constant when estimating the regression of the final model. The constant is not suppressed in the models presented at the workshop.

Source: Sumicsid (2019), 'Project TCB18 Individual Benchmarking Report GTS – 209', July, Table 3.1.

4.1.3 Critique and proposed solution

Sumicsid's cost driver analysis is never outlined in a clear procedure, which is another example of Sumicsid's falling short of good practice in terms of transparency. It is not explained how the final model was selected, what other

[107] Sumicsid (2019), 'CEER-TCB18 project Model Specification Model Results', April, slide 62.

cost drivers were tested and for what reason they were discarded. We note that good-practice model development involves: (i) a selection of candidate cost drivers[108] based on operational insight and interactions with the industry; (ii) a detailed statistical analysis of the candidate cost drivers to identify the most appropriate to include in the final model; and (iii) a robust sensitivity analysis with respect to the choice of cost drivers and modelling assumptions. It is not clear whether Sumicsid followed this process.

Nevertheless, the methods Sumicsid states it has used in model selection (according to the workshop slides and the final report) reveal several concerning aspects, outlined below.

The econometric results are in violation of operational intuition

As can be seen in Table 4.1, in Sumicsid's final benchmarking model two of the four cost drivers used have a *negative* sign.[109] That is, an *increase* in compressor power or weighted pipeline length is associated with *lower* costs in the sample (all else equal). This is counterintuitive and not aligned with operational rationale. There is no reason to assume that a TSO would face lower costs after installing additional compressor power. Sumicsid itself states that 'a DEA model needs outputs that increase cost (input)',[110] yet it has still proposed a model where there is a negative relationship between costs and some outputs.

In other words, while the data tells us that a higher value of these two cost drivers is associated with lower costs, all else equal, DEA will still consider these cost drivers to increase costs.

That coefficients on cost drivers should be of the right sign in an econometric analysis[111] is generally seen as one of the basic requirements in cost driver analysis, including by Sumicsid in the documentation to TCB18[112] and many other assignments.[113]

It is not clear if Sumicsid tried to address this obvious inconsistency. In fact, this is not even acknowledged in the main report at all. If the model selected by Sumicsid does not satisfy this requirement of model development then it cannot be used as a benchmarking model.

The entire model development has to be revisited and a properly defined procedure needs to be put in place. The current benchmarking model was clearly not validated and should not be used to estimate efficiency scores.

The use of ROLS

[108] We note that, in previous international benchmarking exercises, a more detailed description of the candidate cost drivers was provided by the consultants involved. See Frontier Economics, Consentec (2016), 'Gas TSO efficiency analysis for the Dutch transmission system operator', January, section 6.2.3.
[109] Sumicsid (2019), 'Project TCB18 - Individual Benchmarking Report - GTS – 209', July, Table 3.1.
[110] Sumicsid (2019), 'CEER-TCB18 project - Model Specification - Model Results', April, slide 96.
[111] We note that a negative estimated coefficient in an econometric model may be driven by multicollinearity, which is a statistical issue that is not directly relevant in a DEA context. Nonetheless, a negative coefficient could indicate that the characteristic that a cost driver is intended to capture is already captured by the other cost drivers included in the model (at least on average). In this instance, alternative, more appropriate cost drivers could be omitted and the specification must be re-examined. Furthermore, the impact of alternative cost driver specifications must be assessed in a DEA context.
[112] Sumicsid (2019), 'CEER-TCB18 project Model Specification Model Results', April, slide 42–45; and Sumicsid (2018), 'CEER-TCB18 project Methodological Approach', January, slide 26.
[113] See, for example, Sumicsid and Swiss Economics (2019), 'Effizienzvergleich Verteilernetzbetreiber Strom der dritten Regulierungsperiode', April, p. 59; and Sumicsid and Swiss Economics (2015), 'Project E2gas - Benchmarking European Gas Transmission System Operators', June, p. 3.

Sumicsid's use of ROLS appears inappropriate, especially when only the results from this estimation technique are presented. The ROLS estimator explicitly gives less weight to observations that are further from the regression line.

Sumicsid states that the model fits the data well, but it has not presented evidence regarding the number of outliers in the model. We note that outliers are likely to be detected by the ROLS estimator in the final model, as all of the model specifications presented at the final workshop detected several outliers.[114] These outliers which do not appear to fit the model should **not** be ignored, as Sumicsid has done through its use of ROLS, but should be investigated. Sumicsid should have explored why these TSOs do not fit the model well (e.g. they may face different cost drivers, have different operating models or have large data errors) and adjust the model or data accordingly. If Sumicsid deems that these TSOs cannot be made comparable, they should be removed from the sample altogether.

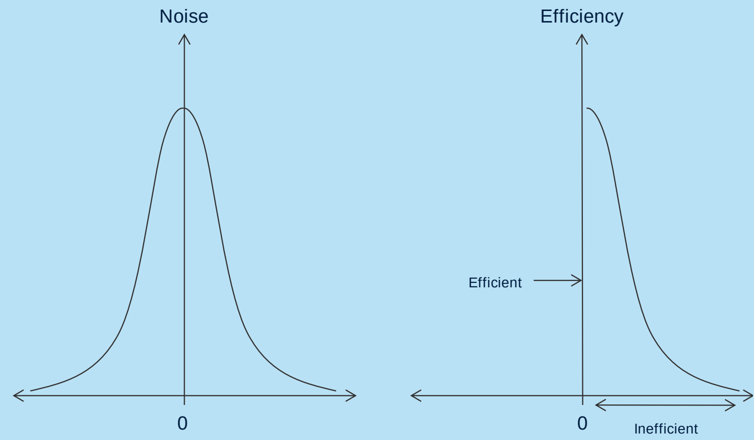
More fundamentally, both the OLS and ROLS used by Sumicsid for cost driver analysis assume that the error term is symmetrical and normally distributed. However, if Sumicsid expects that some TSOs are operating inefficiently, the error term will be skewed, as shown in Box 4.1. Any statistical inference that Sumicsid may have performed in selecting cost drivers is therefore inconclusive.

As a result of these modelling flaws, it is unlikely that Sumicsid's model development procedure has led to an appropriate final model from which unbiased efficiency scores could be estimated. To develop a robust model, Sumicsid should clearly outline its process and candidate cost drivers to the project participants. Participants need to have the opportunity to critique the approach and the candidate cost drivers and should be able to suggest alternatives that need to be considered. Even without this interaction, Sumicsid should have analysed these issues and clearly presented why these issues are not relevant on its final model.

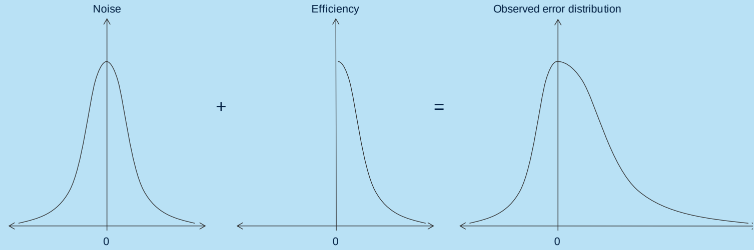
Box 4.1 Statistical inference in the presence of inefficiency

If inefficiency is present in the sample, the residual term of an OLS regression represents two effects. The first effect is pure, normally distributed statistical noise. Statistical noise can have a positive or negative impact on TOTEX, but the mean and median residual will equal zero. The second effect is the inefficiency effect. A TSO's inefficiency will equal zero if it is fully efficient and will be greater than zero if it is inefficient. By definition, a firm cannot be more than fully efficient and the distribution of inefficiency is therefore one-sided, reflecting the higher costs of an inefficient firm relative to a fully efficient firm. A comparison of the distribution of the statistical noise and inefficiency components of the error term is shown in the figure below.

[114] Sumicsid (2019), 'Model Specification Model Results', April, slides 91–95.



The total error term therefore represents a combination (i.e. sum) of these two components and will be asymmetric, as shown in the figure below.



Source: Oxera.

R-squared is not an informative measure of model quality

Sumicsid emphasises the adjusted R-squared as a measure of model quality in the workshops and the final report.[115]

However, R-squared is entirely based on model fit; it does not take into account whether the model is supported by operational intuition (which was clearly violated in TCB18 gas) and thus cannot be used as a method for model selection without being supported by an examination of the regression coefficients. For example, the R-squared is indifferent between estimating a positive or negative coefficient in the model, even though there is an ex ante expectation that all of the cost drivers should have a positive relationship with expenditure.

Concerningly, the R-squared reported by Sumicsid appears to be inflated. This is for the following reasons.

- In its final outputs, Sumicsid has set the intercept to zero manually (see Table 4.1). Sumicsid has not justified why it has forced the intercept in the

[115] Sumicsid (2019), 'Pan-European cost-efficiency benchmark for gas transmission system operators main report', July, Table 5-4; and Sumicsid (2019), 'CEER-TCB18 project - Model Specification - Model Results', April, slides 63–65.

regression to be equal to zero.[116] The reported R-squared in most statistical packages is inflated if the intercept is suppressed.[117]

- ROLS, the estimator used by Sumicsid, explicitly gives less weight to observations that are further from the regression line. Thus the R-squared value will be higher when ROLS is used, despite the model not offering a better fit to the data as a whole.

As demonstrated, not only does Sumicsid focus on a measure of model quality that is not informative, it also does not follow convention in the application of this measure. In fact its application is not consistent within Sumicsid's own outputs. The reported R-squared for the final model in the individual reports is 0.997,[118] whereas the R-squared value presented in the main report is 0.987.[119] This suggests that different methods were used in calculating the model fit in the final report as opposed to the individual reports.

Lasso regression is not evidence on optimal model size

Sumicsid uses Lasso regression to inform its decision to restrict the model to four cost drivers, excluding factors such as urbanity and service quality provisions. Lasso regression is an algorithmic process for model reduction. Starting from a model with a large number of possible cost drivers, the drivers are removed if they do not contribute sufficiently to the model fit. The details of the method are explained in Box 4.2.

[116] Supressing the intercept assumes that there are no 'fixed costs' to the operations of the transmission network, which in turn assumes that there are constant returns to scale. This is inconsistent with the non-decreasing returns to scale assumption applied by Sumicsid in its estimation of efficiency scores in the DEA model.
[117] This is due to the fact that if the intercept is not set to zero, the intercept contributes to the explanation of the variation in the dependent variable, but this is not credited in the R-squared value.
[118] Sumicsid (2019), 'Project TCB18 - Individual Benchmarking Report - GTS – 209', July, Table 3-1.
[119] Sumicsid (2019), 'Pan-European cost-efficiency benchmark for gas transmission system operators main report', July, Table 5-4.

Box 4.2 An overview of Lasso regression

Lasso regression is a type of linear regression aimed at reducing model size (i.e. the number of cost drivers). Lasso regression introduces a penalty term for non-zero parameter estimates, which causes cost drivers with small or statistically insignificant coefficients to be set to zero (i.e. excluded from the model). The regression standard model is set up as follows.

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_n x_{ni} + \varepsilon_i$$

where

- y_i is the TOTEX of TSO i;
- β_0 is the intercept, which represents the value of y when all cost drivers are zero;
- β_j is the cost impact of cost driver j;
- x_{ji} is the value of cost driver j for TSO i;
- ε_i is the random error (residual) for TSO i.

OLS estimates the parameters by minimising the sum of squared residuals across TSOs, as follows.

$$\min \sum \varepsilon_i^2$$

Lasso regression extends the usual OLS minimisation problem to include a penalty term for non-zero values of β_j.

$$\min \sum \varepsilon_i^2 + \lambda \sum |\beta_j|$$

where:

- λ is the degree to which the model is penalised for non-zero values of β_j;
- $|\beta_j|$ is the absolute value of the coefficient on cost driver j.

A larger penalty term leads to more cost drivers being excluded from the model, relative to OLS regression. The size of the penalty term is typically chosen through a cross-validation procedure (e.g. minimising the mean squared error in a validation dataset). If the penalty term is set to zero, the Lasso regression is equivalent to OLS and no cost drivers are excluded from the model.

Source: Oxera.

Importantly, Lasso cannot be used to make a *general* observation about optimum model size for a dataset. Rather, Lasso works with a *specific* set of cost drivers (eleven unnamed ones in Sumicsid's application) and selects a *specific* set of drivers as the most important ones based on a validation criterion (four in Sumicsid's application).

Lasso is specific to the cost drivers, data and functional form used. Once a model has been selected with Lasso, no parameter can be replaced without re-doing the Lasso regression. However, Sumicsid replaced several cost drivers after performing Lasso, without re-examining the results. According to the documentation of Workshop 4, the Lasso selected a model including an output for population density (yArea.Builtup),[120] yet the subsequent models presented do not include this parameter in any form. Conversely, the final model includes the sum of connections, a variable which was not selected by the Lasso. If the results from Lasso regression are to be used, *all* of the outputs of the approach should be considered in the model development, rather than focusing on a limited subset of relevant outputs (as Sumicsid has done). Nonetheless, the application of Lasso in the current case does not constitute evidence on model size.

[120] Sumicsid (2019), 'CEER-TCB18 project - Model Specification – GAS', April, slide 53.

In addition, some of Sumicsid's other approaches to determining the appropriate model size come to a different conclusion. In discussing the appropriate model size in TCB18 electricity, Sumicsid stated that, according to convention, the number of input variables plus the number of output variables in a model should be less than one third of the number of observations. Sumicsid used this as justification for restricting its model to only three outputs.[121] However, if the same logic is applied to TCB18 gas, Sumicsid's model is underspecified and can accommodate up to four additional cost drivers.[122] It is unclear why Sumicsid considered this convention in the electricity study, but not the gas study.

We note that neither Lasso regression nor this convention is an appropriate tool to identify the optimal model size—failure to account for any material driver of expenditure will bias TSOs' efficiency scores, regardless of the size of the model.[123] Sumicsid has not provided sufficient evidence to demonstrate that no material drivers of expenditure have been omitted from its final model.[124]

4.2 The use of NormGrid in benchmarking

4.2.1 Description of the issue

The TCB18 benchmarking exercise uses NormGrid as the primary output variable. As NormGrid is constructed as a weighted sum of different asset classes, it is necessary that the weights on each asset are justified. This is particularly important if TSOs deploy assets in very different ratios to each other.[125] If the weights are inappropriate, a derived variable such as NormGrid may mistake heterogeneity in operating characteristics for inefficiency.

In addition, it needs to be recognised that the weights used to derive NormGrid for an asset (before aggregation) are also *estimated* parameters based on a simplified model of production. It is clearly not the case that all relevant drivers of asset construction and maintenance costs are used when estimating these weights. In addition to this modelling uncertainty, there is the additional uncertainty relating to random data errors and random fluctuations in construction costs. This uncertainty must be accounted for when using NormGrid in the analysis.

4.2.2 Sumicsid's approach

Sumicsid presents the analysis of NormGrid weights in the appendices to its main report.[126] Numerous references are made to the figures being estimates (which are measured with uncertainty) rather than fixed parameters (which are assumed to be measured with negligible uncertainty), but no uncertainty margin is ever calculated.

[121] In the electricity study, there was one input variable (TOTEX) and three outputs, and a total of 17 TSOs. As (1+3)=4 and 17/3~5.7, Sumicsid states that the model is sufficiently discriminatory.
[122] There are 29 TSOs in the gas study, and the model contains one input (TOTEX) and four outputs. As (1+4)=5 and 29/3~9.7, the model can accommodate up to four additional cost drivers (or input variables) and still be sufficiently discriminatory (according to the convention Sumicsid has appealed to).
[123] The limitations of approaches used to restrict model size are discussed in Deuchert, E. and Parthasarathy, S. (2018), 'Gibt es eine »optimale Modellgröße«?', December.
[124] Examples of omitted environmental factors are outlined in section 4.3 and the flaws in Sumicsid's second-stage analysis are outlined in section 5.4.
[125] We note that, in the current sample of gas TSOs, there is likely to be a significant heterogeneity in asset structure given the heterogeneity in activities performed by the TSOs. For example, TSOs that transport gas across vast distances are likely to have more pipelines, whereas TSOs that perform activities closer to that of a DSO will have more compressors.
[126] Sumicsid (2019), 'Norm Grid Development - TCB18 PROJECT', February.

Sumicsid then chooses the weights on assets based on an average cost estimate of each asset.[127] In the appendices to its main report, Sumicsid states:

> The calibration of the asset weight systems is made through linear regression towards the Capex and Opex data obtained in the project. This step scales the relative NormGrid metric towards average practice (not best practice) such that the relevant cost measures are attributed to the size proxy.

The results of this regression are not part of the final report or its appendices and there appears to have been a separate treatment of OPEX and CAPEX weights. Specifically:

- CAPEX weights appear to have been undertaken by asset category, as the groups have different weights;
- OPEX weights seem to have been estimated on an overall basis rather than by asset category.[128]

4.2.3 Critique and proposed solutions

The calculation of NormGrid involves estimates and assumptions at many stages. For instance, in the calculation of the NormGrid value for pipes, Sumicsid states:

> The average unit cost [of internal coating] is **estimated** to 10 € / m2.[129]
>
> In the cost estimation of pipes, we will **assume** the average distribution of the following class locations[130]
>
> [emphasis added]

The validation exercise for pipelines on ACER data that Sumicsid considered shows that there is considerable variation in costs when controlling for Sumicsid's parameters.[131] The 'noise' in these estimates and assumptions made are not addressed at any point in the analysis. DEA (as applied by Sumicsid) is not able to deal with this noise in the data at all. This will add uncertainty to the estimated efficiencies, as TSOs on the frontier may have benefited from this effect while inefficient TSOs may be penalised.

This uncertainty should have been robustly addressed by Sumicsid. Ideally, a margin of uncertainty should have been included by Sumicsid, along with every assumption and estimation made in the development of NormGrid. This would then enable the calculation of an aggregate uncertainty margin that could have been applied to the NormGrid estimate of each asset. However, these estimates are not available and cannot be derived from the material provided by Sumicsid.

In addition to this uncertainty in the NormGrid estimate for individual assets, our review of TCB18 electricity also revealed two observations that cast doubt on Sumicsid's approach to aggregating across asset types.

- The electricity TSOs are very heterogenous in the assets they deploy to meet their service obligations.[132] For this reason, a TSO's performance on the NormGrid measure is likely to be sensitive to the relative weight on each

[127] Sumicsid (2019), 'Pan-European cost-efficiency benchmark for gas transmission system operators main report', July, p. 26.
[128] All assets have the same OPEX weight.
[129] Sumicsid (2019), 'Norm Grid Development - TCB18 PROJECT', February, section 2.1.2.2.
[130] Sumicsid (2019), 'Norm Grid Development - TCB18 PROJECT', February, section 2.1.1.
[131] Sumicsid (2019), 'Norm Grid Development - TCB18 PROJECT', February, Figure 2-3.
[132] Oxera (2020), 'A critical assessment of TCB18 electricity', April, Figure 4.4.

asset.[133] It is likely that gas TSOs are also heterogenous in assets, given the differences in their service task.

- We could not validate the weights on each NormGrid measure with regression analysis. That is, using the same approach that Sumicsid describes in its appendices, we derive different weights on each asset class from those presented by Sumicsid in its outputs.

These observations indicate that, in electricity, there is a high level of uncertainty regarding the appropriate weights on the asset classes, and TSOs' relative performance on the NormGrid measure will be sensitive to the exact values of the weights. As the overall approach to NormGrid estimation is the same in TCB18 gas, it is likely that a similar issue arises. This uncertainty should have been modelled by Sumicsid, instead of relying on a point estimate.

We note that the average unit cost (defined as TOTEX divided by NormGrid) should be relatively close to one if NormGrid is a measure of 'average practice'. However, the average unit cost is about 1.21,[134] indicating that the NormGrid measure may not be a good proxy for average practice.

Moreover, it has been argued that the use of the cost weights for the aggregation of the physical assets contradicted the principle of DEA, which chooses input and output weights in such a way as to give the firm the highest efficiency score possible.[135] For these reasons, Sumicsid should have considered controlling for each asset class as separate outputs. This has the benefit that TSOs will only be compared with peers that have similar asset bases. Indeed, according to Sumicsid's own conventions,[136] the model can easily accommodate more output variables, so each asset class can be controlled for a separate output without affecting the discriminatory power of the model.

In the meeting with the ACM, Sumicsid argued against this approach as it would amount to giving each asset class an arbitrary weight and TSOs may be estimated to be fully efficient purely because they deploy an unusual mix of assets and therefore have no comparators. However, it may be appropriate to only compare TSOs that have similar asset mixes, as the solutions available to TSOs with a disproportionate amount of pipes (for example) may be very different from the solutions available to TSOs with a disproportionate amount of compressors. Moreover, if Sumicsid is concerned that TSOs may give excessive weight to certain assets and no weight to others, it is possible to impose weight restrictions on the model (although the weight restrictions must be robustly validated).

4.3 Environmental factors

4.3.1 Description of the issue

As a TSO's main task consists of transporting energy across a country, the features of that country's environment, such as land use, climate and

[133] Note that most TSOs attached no or minimal weight to the NormGrid output in Sumicsid's model. As such, changing the weight on NormGrid assets had a minimal impact on the estimated efficiency score of each TSO. By 'performance on the NormGrid measure' we mean unit cost performance. See Oxera (2020), 'A critical assessment of TCB18 electricity', April, Figure 5.4.
[134] The mean value of TOTEX is 151,486,489 and the mean value of NormGrid (adjusted for slope) is 125,349,439. The ratio of the two is 1.21. See Sumicsid (2019), 'Pan-European cost-efficiency benchmark for gas transmission system operators main report', July, table 5-3.
[135] Brophy Haney, A. and Pollitt, M.G. (2013), 'International benchmarking of electricity transmission by regulators: A contrast between theory and practice?', *Energy Policy*, **62**, November, pp. 267–81.
[136] Note that we do not consider these conventions to be appropriate. However, this does highlight the inconsistency in Sumicsid's analysis.

topography, can be a significant driver of their costs. Costs associated with environmental factors can take the form of access costs, increased maintenance costs and increased costs of routing pipelines around obstacles.

Although TSOs should act in such a manner as to mitigate the impact of these environmental costs (e.g. through grid planning), these environmental characteristics are not controllable by management (i.e. they are exogenous) and thus need to be accounted for in the benchmarking process, either through the model directly or through post-modelling adjustments.

4.3.2 Sumicsid's approach

Sumicsid controls for two types of environmental conditions by multiplying cost drivers by an environmental complexity factor based on the characteristics of a TSO's service area.[137] Specifically, Sumicsid controls for:

- slope complexity, by multiplying NormGrid by a slope complexity factor;
- humidity, by multiplying total pipe length by a humidity complexity factor.

The weights used to construct some of the environmental adjustments are not discussed or presented in the final report or its appendices,[138] but are presented in one of the workshops.[139] It is not clear on what basis these weights are derived.

Sumicsid also makes several statements regarding the correlation between these specific environmental factors and other environmental factors such as land use (e.g. population density) and sub-surface features (e.g. gravel), thus implying that overall environmental conditions are captured in its model.

> Extensive statistical tests revealed correlations and interaction between several of the factors, e.g. vegetation and land use type, subsurface features and topography […][140]
>
> Population density best covered by land use factors (incl. GIS-level density areas). NG_Area is 99% correlated with NG_Slope for gas, leading to a choice where Slope makes a stronger technoeconomic sense. [141]

4.3.3 Critique and proposed solution

Sumicsid presents no evidence that it has validated the calculated environmental factors on the TCB18 data. This may lead to a situation where the corrections are not appropriate for the sample. In TCB18 electricity, for instance, we found that the environmental adjustments that Sumicsid made in its model had a negative correlation with costs per unit of NormGrid.[142] That is, TSOs that operate in areas that should be more costly to operate in (based on land use factors) have lower unit costs than TSOs that operate in areas that should be less costly. This suggests that Sumicsid did not appropriately examine the correlations in the data and the relevance of these environmental factors in TCB18. As this validation exercise was not undertaken by Sumicsid in TCB18 electricity, we consider it unlikely that this was performed in TCB18 gas (and, in any case, it is not present in the outputs that we have reviewed).

[137] It is not clear from the final outputs whether the service area is defined as the area actually served by a TSO, or represents the characteristics of the entire country or region.
[138] The weights for the humidity and gravel complexity factors are presented in the NormGrid appendix.
[139] Sumicsid (2019), 'Model Specification Model Results', April, slides 52 and 54.
[140] Sumicsid (2019), 'Pan-European cost-efficiency benchmark for gas transmission system operators MAIN REPORT', July, para. 5.04.
[141] Sumicsid (2019), 'Pan-European cost-efficiency benchmark for gas transmission system operators MAIN REPORT', July, Table 3-3.
[142] Oxera (2020), 'A critical assessment of TCB18 electricity', April, p. 60.

Beyond this simple validation problem, we have identified three further issues with Sumicsid's environmental adjustment.

- Sumicsid's approach to adjusting output variables by environmental complexity factors is inconsistent with non-parametric analysis and may inadequately account for environmental complexity.
- Sumicsid's environmental adjustment factors appear to ignore asset location.
- Sumicsid has not ensured that all relevant environmental factors are adequately accounted for in its model.

Sumicsid's overall approach to accounting for environmental factors is inappropriate

Adjusting existing output variables for environmental factors imposes a rigid, arbitrary and unexplained relationship between costs and environmental factors. Sumicsid has not provided any evidence (empirical or operational) that such weight restrictions are valid.

A TSO that operates in a complex environment (as measured by the environmental complexity factors) may not benefit from Sumicsid's adjustment under this approach. For example, if a TSO attaches no weight to the NormGrid variable when estimating its efficiency score, it will not benefit from the slope adjustment, regardless of how 'sloped' its operating environment is. Similar arguments could be made for Sumicsid's humidity adjustment to pipeline length. Sumicsid's approach is thus equivalent to imposing extremely strict weight restrictions on the relative importance of slope complexity and NormGrid, and humidity and pipeline length.

Importantly, this is not just a hypothetical argument—in our assessment of TCB18 electricity, we found that for most TSOs NormGrid was not a primary driver of efficiency, and therefore did not benefit from any environmental adjustment.[143]

Additionally, the use of a composite variable may be problematic if the returns to scale are assumed to be non-variable (i.e. constant or increasing). Cost drivers adjusted for environmental factors are composite variables. The composite variable consists of a factor reflecting both environmental complexity (e.g. slope) and a cost driver (e.g. NormGrid). Thus the composite variable has to reflect both a cost driver and terrain. The non-VRS assumptions require that both TOTEX and outputs are scalable by the same scalar in efficient production. Sumicsid has provided neither conceptual reasoning nor empirical evidence that this is an appropriate assumption in this context.

After scaling it is impossible to disentangle the component reflecting terrain and the one reflecting asset base. It is not evident from the DEA model which characteristic is being compared—the asset base or the environmental conditions.[144] This makes it difficult to determine whether the benchmark (i.e.

[143] We understand that, if the slope complexity factor is replaced with a land use complexity factor, GTS's efficiency score does not change. In this instance, it is very likely that GTS does not attach any weight to NormGrid, and therefore does not benefit from any environmental adjustments to NormGrid.
[144] Assume that the original NormGrid of a large TSO was *A* and *jA* following adjustment. Assume the benchmark was derived from a smaller TSO which had unadjusted NormGrid *B,* adjusted by the environmental factor *k*. Thus, the smaller TSO's NormGrid is *kB* following adjustment. The benchmark will be derived by scaling *kB* to *nkB* where n>1, as the TSO is smaller. Moreover, *nkB≥jA*. In *nkB*, we cannot disentangle the component reflecting terrain from the one reflecting asset base.

peer companies) is appropriate and necessitates a detailed peer analysis, which Sumicsid did not undertake.

Sumicsid's environmental adjustment factors ignore asset location.

In the final report, Sumicsid suggests that the granularity of the GIS-based environmental data is 'very good', with cells of a size of $25m^2$ for the slope factor.[145] However, the final report does not mention that Sumicsid did not take advantage of this granularity to estimate TSO-specific environmental factors. Instead, the adjustment factors are calculated by aggregating the data to a country level for most TSOs. For example, we note that National Grid Electricity Transmission Plc. has the same environmental adjustments as National Grid Gas Transmission Plc. despite the electricity TSO serving only England and Wales, while the gas TSO serves the whole of Great Britain (including Scotland, which is served by Scottish Power and SHE-T in electricity). Thus, no account is taken of the actual location of pipelines and other assets.

For instance, [*Vertrouwelijk: a European TSO*] receives a large adjustment in the slope factor of 1.5, as 27.6% of [*Vertrouwelijk: the country*] is considered to be mountainous based on GIS data. However, [*Vertrouwelijk: the European TSO's*] assets are not evenly distributed across the country, but concentrated in the valleys, which are typically flat and easier to operate in than mountainous regions. Therefore, it is likely that the adjustment factor of 1.5 is excessive. Indeed, analysis of GIS data conducted by GTS indicates that [*Vertrouwelijk: the European TSO's*] slope complexity factor would be 1.25 if it was estimated based on the location of [*Vertrouwelijk: the European TSO's*] assets. That is, if the location assets were taken into account, [*Vertrouwelijk: the European TSO's*] adjustment to NormGrid would be 50% lower.[146] This will lead to a bias in the calculated adjustments factors unless the magnitude of the error is the same for all TSOs in the sample.

Another example is the construction of the land use factor, which Sumicsid did not use in TCB18 gas. The land use factor is largely based on the share of area covered in urban fabric and forests. While for forests a similar logic as for mountainous areas prevails (i.e. they are likely to be avoided if construction costs are high), the same cannot be said for urban areas and infrastructure. Clients (which TSOs like GTS are obliged to connect), and thus assets, may in fact be disproportionately likely to be located close to these terrain features. This was not taken into account by Sumicsid when calculating the environmental adjustment factors.

Sumicsid's statements regarding the correlation between its selected environmental factors and other environmental adjustments are misleading

As shown in Table 4.2, the correlation between the different complexity factors in the TCB18 gas sample is limited: only the land use and the humidity categories are more than weakly correlated. Moreover, the slope factor is *negatively* correlated with the humidity and subsoil factors. It is unclear to us how Sumicsid has concluded that the environmental factors are highly

[145] Sumicsid (2019), 'Pan-European cost-efficiency benchmark for gas transmission system operators MAIN REPORT', July, p. 10.
[146] We understand that [*Vertrouwelijk: a European TSO*] is assessed to be an efficient outlier in Sumicsid's model. If TSOs are detected as outliers in the model estimation approach, their data (and the model specification) should be re-visited to identify errors or inconsistencies.

correlated with each other, so Sumicsid needs to validate its statement with specific empirical evidence.

Table 4.2 Correlation between environmental factors

	Slope	Humidity	Land Use	Gravel
Slope	100%	-72%	-27%	21%
Humidity		100%	40%	-25%
Land Use			100%	12%
Gravel				100%

Source: Oxera analysis, based on TCB18 gas data.

In fact, Sumicsid's choices of complexity variables (slope and humidity) may bias the analysis against GTS. In response to a cost adjustment claim brought forward by GTS, arguing for the inclusion of density in the benchmarking, Sumicsid states: [147]

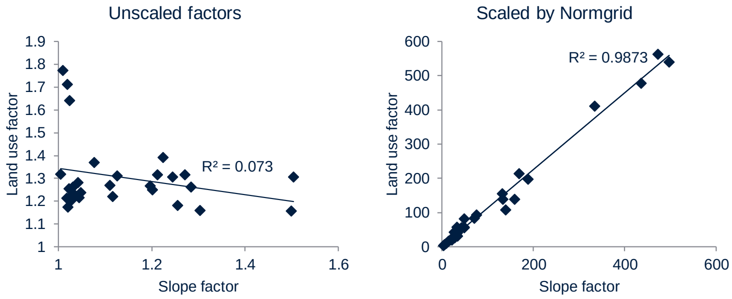
> Population density best covered by land use factors (incl. GIS-level density areas). NG_Area is 99% correlated with NG_Slope for gas, leading to a choice where Slope makes a stronger technoeconomic sense.

This statement appears to be misleading for several reasons.

- First, claiming correlation between environmental factors, based on a measure scaled by NormGrid, is not informative, as the comparison will be dominated by scale. Figure 4.1 provides an illustrative example in which we have plotted the environmental adjustments for slope and land use in TCB18 gas against each other and then scaled them by the NormGrid variable in TCB18 gas. Despite the negative correlation in environmental factors, the scaling by NormGrid will suggest a near-perfect positive correlation. Sumicsid could have used similar evidence (i.e. a correlation between two adjusted NormGrid variables) to justify any arbitrary adjustment to NormGrid.
- Second, the analysis presented in Table 4.2 shows that the land use factor and the slope factor are in fact negatively correlated.
- Third, even the land use factor itself is not primarily determined by density. Our analysis of the land use factor suggests that the share of the adjustment for land use determined by urbanity was less than 25% of the total adjustment on average.

[147] Sumicsid (2019), 'Pan-European cost-efficiency benchmark for gas transmission system operators MAIN REPORT', July, Table 3-3.

Figure 4.1 Correlation between environmental factors



Note: At present, we do not know which value of NormGrid corresponds to which TSO. Exact correspondence is not needed for this example, as scaling dominates regardless of the assignment.

Source: Oxera analysis on the basis of TCB18 data. Relative NormGrid values are sourced from Sumicsid (2019), 'Project TCB18 Individual Benchmarking Report GTS – 2019', July, Figure 4.15.

In general, Sumicsid has presented no evidence to suggest that the impact of population density on costs is robustly accounted for in its model. Indeed, the environmental adjustments are both negatively correlated with population density. As GTS operates in one of the densest environments in the sample, this may result in a significant bias in GTS's estimated efficiency score.

As discussed in section 4.1.3, the model can easily accommodate more output variables using Sumicsid's own conventions.[148] As such, a feasible solution to this issue would be to explicitly control for some aspects of environmental complexity (such as population density) as exogenous drivers of expenditure. Such an approach would be consistent with the non-parametric nature of DEA and allow for a more robust comparison across TSOs. Without the appropriate treatment of environmental factors, Sumicsid cannot reasonably conclude that its model (and the resulting efficiency estimates) is robust.

[148] It is important to note that these conventions are not based on evidence, and the omission of any material driver of expenditure will bias the estimated efficiency scores, regardless of the size of the model.

5 Application and validation

5.1 The use of DEA in Benchmarking

5.1.1 Description of the issue

DEA is one method of assessing the efficiency of TSOs. Although it has certain advantages over other methods (for example, it does not impose strict assumptions on the relationship between inputs and outputs), it suffers from several limitations.

- DEA is a deterministic method of efficiency assessment.[149] In this respect, it treats the data and assumptions 'as given' and makes no allowance for uncertainty in the variables.
- The results are contingent on assumptions imposed on the model (e.g. returns to scale) which have not been sufficiently motivated by Sumicsid.

It is therefore common for regulators to use alternative benchmarking methods, either to directly inform the efficiency target or as a cross-check to the results from DEA. For example, the Bundesnetzagentur uses four models to estimate a DSO's relative efficiency, two of which are estimated via DEA and two of which are estimated via SFA. A DSO's efficiency score is the best of the four, with a lower bound of 60%.

SFA is particularly relevant when we suspect a lot of uncertainty in the data. As highlighted in section 3.1, there is significant uncertainty in the data used by Sumicsid (this includes issues with its data processing and data adjustments), and it is therefore important that the results from DEA are supported by SFA models.

5.1.2 Sumicsid's approach

Sumicsid has not used SFA or any other benchmarking method to validate its model, nor has it used SFA to assess the robustness of the TSOs' efficiency scores.

Sumicsid states:

> In a study of European gas TSOs, the number of observations is too small for a full-scale application of SFA as main instrument. We have therefore used DEA as our base estimation approach, in line with regulatory best practice and earlier studies such as E2GAS and E3GRID.[150]

Note that Sumicsid used SFA to validate its DEA model in e2gas,[151] despite having a smaller dataset.[152]

5.1.3 Critique and proposed solution

All models are imperfect representations of reality, and any one model could overestimate or underestimate an individual TSO's efficiency.[153] As such, it is

[149] Note that there are several approaches in the DEA literature that aim to mitigate the deterministic nature of the DEA (e.g. stochastic DEA, StoNED). However these are not used by Sumicsid.
[150] Sumicsid (2019), 'Pan-European cost-efficiency benchmark for gas transmission system operators main report', July, p. 30.
[151] Sumicsid, Swiss Economics (2016), 'Project E2GAS Benchmarking European Gas Transmission System Operators', June, pp. 44–45.
[152] In e2gas, Sumicsid used one year of data for 22 TSOs to estimate its model. In TCB18, Sumicsid had access to data for 29 TSOs in at least one year, leading to a total sample size of 70 observations.
[153] Indeed, this point was accepted by Sumicsid and the ACM during the meeting with Oxera.

important to consider valid alternatives, both in terms of cost driver selection and estimation technique.

While most empirical investigations (including both DEA and SFA) perform better on larger samples, there is no fixed rule as to how many observations a model needs. Indeed, SFA and other econometric methods have been used on smaller sample sizes than that available to Sumicsid,[154] including by Sumicsid itself in previous iterations of the gas study. As such, sample size in and of itself is not a valid justification for ignoring SFA, and the appropriateness of a method has to be determined empirically on the data and model used.

Table 5.1 highlights the inconsistencies between e2gas and TCB18 on the topic. As can be seen in the table, the text regarding the appropriate benchmarking model is nearly identical in the two reports. However, Sumicsid has removed its text from the e2gas report on the use of SFA (emphasised in bold) for the TCB18 report.

Table 5.1 Sumicsid's inconsistent view on SFA

E2GAS section 4.82-4.84	TCB18 section 4.91-4.92
Econometrics has provided a portfolio of techniques to estimate the cost models for networks, illustrated in Table 4-2 below. Depending on the assumption regarding the data generating process, we divide the techniques in deterministic and stochastic, and further depending on the functional form into parametric and non-parametric techniques. These techniques are usually considered state of the art and are advocated in regulatory applications provided sufficient data is available.	Econometrics has provided a portfolio of techniques to estimate the cost models for networks, illustrated in Table 4-4 below. Depending on the assumption regarding the data generating process, we divide the techniques in deterministic and stochastic, and further depending on the functional form into parametric and non-parametric techniques. These techniques are usually considered state of the art and are advocated in regulatory applications provided sufficient data is available.
[Table 4-2 omitted]	[Table 4-4 omitted]
In a study of European gas TSOs, the number of observations is too small for a full scale application of SFA as main instrument. We have therefore used DEA as our base estimation approach. **As part of the robust check, we have additionally estimated the same model using SFA. Part of the motivation for this is also to discipline the modelling effort. In a good model specification, our experience is that the DEA and SFA approaches lead to comparable results, i.e. the average efficiencies should not deviate too much and the correlation of DEA and SFA efficiencies should be reasonably high.** [emphasis added]	In a study of European gas TSOs, the number of observations is too small for a full-scale application of SFA as main instrument. We have therefore used DEA as our base estimation approach, in line with regulatory best practice and earlier studies such as E2GAS and E3GRID. The DEA method is by now well established in the scientific literature as well as in regulatory applications, and we shall therefore not provide a theoretical description of it here
Benchmarking methods like DEA **and SFA** are by now well established in the scientific literature as well as in regulatory applications, and we shall therefore not provide a	

[154] For example, the ORR used estimated SFA models with 14 infrastructure managers (although the time series component was longer). The ORR also performed SFA on a sample of 50 observations for its determination of the efficiency on Network Rail as part of the PR18 price control. See The Office of Rail and Road (2013), 'PR13 Efficiency Benchmarkings of Network Rail using LICB', August, p. 6, (2018), 'PR18 Econometric top-down benchmarking of Network Rail A report', July, p. 43.

theoretical outline of these methods.
[emphasis added]

Source: Sumicsid, Swiss Economics (2016), 'Project E2GAS Benchmarking European Gas Transmission System Operators', June, p. 30, Sumicsid (2019), 'Pan-European cost-efficiency benchmark for gas transmission system operators main report', July, p. 30.

It is not clear why SFA was considered a suitable validation approach in e2gas, but is no longer relevant in TCB18, when the sample size has increased significantly between the two iterations of the study. Indeed, in some workshop material presented in the early stages of the TCB18 project, Sumicsid states:

> SFA and other parametric methods may give additional info and serve as a crosscheck for DEA[155]

Sumicsid gives no justification for omitting the results of SFA modelling from its final outputs. If SFA models gave counterintuitive results (e.g. negative coefficients on some cost drivers), failed to converge[156] or did not detect statistically significant inefficiency,[157] then this is still a useful result for TSOs and NRAs in interpreting the results from the DEA model that are presented in the TCB18 outputs. In fact, such a result is an essential caveat to Sumicsid's core analysis. Indeed, in our review of TCB18 electricity, we found that the model did not detect any statistically significant inefficiency across several different modelling assumptions,[158] and this severely limits the interpretability of the efficiency scores estimated by Sumicsid's DEA model.

As a general related observation, we note that despite deriving its model on a panel dataset (i.e. data over time across TSOs), Sumicsid has not effectively used all the information it has at its disposal. Instead, it has focused on a single year's data without justification. Also, Sumicsid should have validated the outputs from DEA (e.g. peers and weights) to show that the method was appropriate for the dataset, but it has not.

The non-consideration of SFA is also in conflict with the lack of consideration given to sample size in Sumicsid's current approach—for example, with respect to its second stage analysis and the dominance test, which rely on large samples (among other conditions) to be appropriate. Sumicsid made no attempt to justify the size of its sample for this analysis and has not validated the DEA output to show that it is valid for the dataset, yet it has relied on sample size alone to justify ignoring SFA models from the evidence base.

5.2 Returns to scale

5.2.1 Description of the issue

'Returns to scale' relates to how changes in inputs (i.e. TOTEX) are linked to changes in outputs (e.g. NormGrid) for efficient companies.

The choice could be a matter of policy in a national benchmarking exercise. For example, if a regulator wishes to encourage firms to move to a more productive scale size (e.g. through mergers with other firms), it could use a CRS assumption. This policy perspective, however, is not applicable in the TCB18 assessment, which is a cross-country comparison. The scale of a TSO,

[155] Sumicsid (2018), 'CEER-TCB18 project Methodological Approach', January, slide 22.
[156] If a model does not converge, it means that the iterative procedure used to estimate the model results in an endless loop. In this sense, the model cannot be estimated. This could be because of poor starting values, wrong distributional assumptions or general model misspecification.
[157] In our review of TCB18 electricity, we found that the SFA model could not detect any statistically significant inefficiency in the sample under a many different parametric assumptions.
[158] Oxera (2020), 'A critical review of TCB18 electricity', April, section 5.5.3.

particularly those that cover their entire country, is not within management control.[159] As such, a VRS assumption may be more appropriate. In any case, the returns-to-scale assumption—as with any other assumption made in the modelling—should be empirically validated.

5.2.2 Sumicsid's approach

Sumicsid uses a non-decreasing returns to scale (NDRS) assumption. Sumicsid states that this is supported by the following statistical evidence.

- **A Banker F-test for returns to scale on the DEA efficiencies.** In this test, the efficiencies estimated under CRS are compared to efficiencies under alternative RTS assumptions. The test statistic is similar to that used in the dominance test, and compared to an F-distribution. If the test statistic is statistically significant, this is taken as evidence for the alternative assumption.
- **The sum of coefficients in a log-linear regression.** If the sum of coefficients is less than one, adding 1% to every output increases costs by less than 1%, indicating increasing returns to scale.[160]

Sumicsid does not present the results of these tests in its final outputs for TCB18. This is in contrast to Sumicsid's approach in e2gas, where Sumicsid did present the outputs for some statistical tests regarding returns to scale.[161]

Sumicsid further states that the NDRS assumption is supported by operational intuition.

> This [Sumicsid's NDRS returns-to-scale assumption] is also conceptually appealing. A TSO can be small due to the size of the country or by the service area it has to serve and there may be an element of fixed costs involved in the operation of any TSO. On the other hand, if a TSO is suffering from extra cost of being large, it is likely that a reorganization of the TSO to imitate a combination of smaller TSOs could improve cost efficiency.[162]

5.2.3 Critique and proposed solution

In our critique of TCB18 electricity, we performed the same tests as Sumicsid and found no evidence supporting an NDRS assumption.[163] Indeed, most tests supported either a DRS, CRS or VRS assumption, counter to Sumicsid's statements. Although we cannot perform the same validation exercise in TCB18 gas due to data limitations, we note that the following observations provide contradictory evidence regarding Sumicsid's claims on the appropriate returns-to-scale assumption.

- Sumicsid forces the intercept to be zero when presenting its final model in the individual reports. This implicitly assumes that there are no 'fixed

[159] For example, in addition to TSOs that may be unable to increase their scale due to national boundaries (or regulatory imposed boundaries), it may not be feasible for a TSO to reduce its size if regulatory restrictions are in place. We note that the German TSOs are better able to increase or decrease the scale of their operations through mergers and demerges than national TSOs, such as GTS.
[160] Sumicsid (2019), 'Pan-European cost-efficiency benchmark for gas transmission system operators MAIN REPORT', July, p. 36.
[161] For example, see Sumicsid, Swiss Economics (2016), 'Project E2GAS Benchmarking European Gas Transmission System Operators', June, Figure 5-3.
[162] Sumicsid (2019), 'Pan-European cost-efficiency benchmark for gas transmission system operators MAIN REPORT', July, p. 37.
[163] Oxera (2020), 'A critical assessment of TCB18 electricity', Prepared on behalf of European electricity TSOs, April, section 5.1.3.

costs'[164] to production, which indicates that the technology is CRS (see section 4.1.3).

- Of the five models presented in the final workshop, only one has a positive intercept (albeit statistically insignificant) and the remaining four have negative intercepts (one of which is statistically significant).[165] The evidence presented in this workshop therefore indicates that either a DRS, CRS or VRS assumption would be appropriate.

The empirical evidence published in TCB18 suggests that the statistical support for the NDRS assumption is, at best, weak. Furthermore, there are many alternative statistical tests that Sumicsid can and should have considered as evidence to support its returns-to-scale assumption.

- **Bootstrap-based tests for model fit.** The Banker test for returns-to-scale requires large and independent samples and strong parametric assumptions regarding the distribution of inefficiency.[166] Sumicsid has not validated whether these requirements are met in this instance. Indeed, the samples are not independent by construction. An alternative approach could be a perform a bootstrap-based test for model fit, similar to the alternative outlier tests outlined in section 5.3.[167] Like the Banker Test, it rests on testing for statistical significance of the mean efficiencies under CRS versus VRS assumptions.
- **Translog models.** The parametric tests considered by Sumicsid can only test whether the data exhibits decreasing, increasing or constant returns to scale (depending on whether the sum of coefficients is greater than, less than or equal to one). Therefore it cannot test whether a VRS assumption is appropriate. To do so, one must consider more flexible functional forms that allow the cost-impact of scale (i.e. the returns to scale) to vary with a TSO's size by including higher order terms in the regression model.
- **SFA models.** As discussed in section 4.1.3, statistical inference based on OLS-type estimators is invalid in the presence of inefficiency due to the asymmetric nature of the error term. For this reason, it is appropriate to consider estimation approaches that can explicitly account for inefficiency, such as SFA.

Sumicsid's assertion that a large TSO can reorganise to imitate a combination of smaller TSOs is not supported by the evidence. The policy recommendations from such an approach is, if a TSO is too large, then it should be separated into smaller operating networks (or management units).

For the productivity gains to be made either the large TSO needs to be reduced to the most productive scale size *for its mix of outputs*, or split up into several TSOs, each one enjoying the most productive scale size. Even ignoring the substantial reorganisation costs involved in this, neither approach is feasible in practice. For instance, GTS cannot reduce the scale of operations, as it is obliged to connect clients in the Netherlands and is prohibited from separating into two or more smaller TSOs. Moreover, breaking up the large TSO cannot guarantee that any one of the resulting smaller TSOs

[164] The intercept represents the value of the dependent variable (in this case, TOTEX) when all of the cost drivers are zero. In other words, it is the costs of 'production' when the TSO produces no outputs, and is therefore interpreted to be the fixed costs of production.
[165] Sumicsid (2019), 'CEER-TCB18 project Model Specification Model Results',
[166] The limitations of the Banker test are discussed in more detail in section 5.3.
[167] The approach is based on drawing repeated subsamples from the data. Simar, L. and Wilson, P.W. (2011), Inference by the m out of n bootstrap in nonparametric frontier models, *Journal of Productivity Analysis*, **36**, pp. 33–53.

will necessarily have the same mix of outputs (terrain, NormGrid, connections etc.) as the large TSO but at the most productive scale size. In fact it is possible that any feasible subdivision of the TSO will itself operate at a non-optimal scale.

If the regulator has an explicit objective for companies to operate at the optimal scale size (e.g. through mergers and demergers) then it may be appropriate to assume a CRS technology in the benchmarking model, even if the 'true' technology is VRS.[168] For example, the Norwegian energy regulatory framework ensures that the distribution system operators (DSOs) are incentivised to undertake structural reforms and are compensated if their efficiency position in the benchmarking model worsens due to any such reform. This policy perspective, however, is not applicable in the TCB18 assessment, which is a cross-country comparison. The scale of a TSO, particularly those that cover their entire country, is not within management control.[169] Thus, if the true technology is VRS,[170] there is no valid reason to assume any other technology.

Moreover, if Sumicsid insists on using an NDRS assumption, it must address the question of why large TSOs are systematically estimated to be inefficient (which would cause the statistical tests outlined above to conclude that a VRS technology is appropriate). One hypothesis (implicitly assumed by Sumicsid) is that larger TSOs happen to be inefficient by chance in this sample, and that a larger or different sample would show a different relationship between size and productivity. However, this cannot be assumed without evidence, as Sumicsid has done. An alternative hypothesis could be that Sumicsid has omitted key drivers of expenditure that (e.g. by chance) happen to be correlated with scale in the sample, causing a bias in the estimated efficiency scores of large TSOs. For example, several of the largest TSOs in both gas and electricity happen to operate in relatively densely populated countries (NGET and TenneT in electricity, NGGT and GTS in gas). Additionally, it may be the case that TSOs that have accumulated a number of assets over time (thereby being 'large') now operate a relatively old grid. As Sumicsid's model does not control for asset age, this could bias the analysis against large TSOs.[171]

In any case, if the observed relationship between cost and output is not consistent with Sumicsid's operational expectations (that are themselves not justified in its report), Sumicsid should have investigated and documented the causes of such inconsistencies instead of ignoring the problem (as it has done in this case).

This issue is particularly relevant for GTS as it is the second-largest TSO in the sample, as measured by its value for NormGrid.[172] Indeed, the scale efficiency analysis that Sumicsid presents in the individual reports appears to imply that GTS would be estimated to be fully efficient under a variable returns to scale

[168] This is for example the case in German DSO regulation where BNetzA employs a CRS technology assumption with the objective of encouraging structural reforms to achieve efficient scale. We would note that there are conceptual and empirical inconsistencies in the application of this this policy decision of Bundesnetzagentur by its consultants.
[169] For example, in addition to TSOs that may be unable to increase their scale due to national boundaries (or regulatory imposed boundaries), it may not be feasible for a TSO to reduce its size if regulatory restrictions are in place.
[170] VRS may be present in the model if (i) there are genuine diseconomies of scale in the organisation of TSOs or (ii) Sumicsid's model is mis-specified and omits important drivers of cost for the large TSOs in the sample. For a discussion on the latter argument, see Banker (2016), 'The Returns to Scale Assumption in Incentive Regulation', March.
[171] In work for TenneT, we found that omitting the impact of asset age on costs led to an underestimation of TenneT's efficiency. See Oxera (2020), 'Analysis of TenneT's estimated efficiency under TCB18', August, section 3.
[172] Sumicsid (2019), 'Project TCB18 - Individual Benchmarking Report - GTS – 209', July, Figure 4.15

(VRS) assumption.[173] The analysis also suggests that GTS is being benchmarked against TSOs that could be significantly different to itself (in terms of size and complexity) insofar that many of the solutions implemented by its peers may not be feasible for GTS to replicate. In other words, given the differences in scale, GTS's peers may not be sufficiently comparable to the GTS to allow for a robust comparison.

5.3 Outlier analysis

5.3.1 Description of the issue

The DEA model, as applied by Sumicsid, determines the frontier based on those TSOs deemed to be efficient without making any allowance for errors (either data errors or model mis-specification). For the TSOs to be sufficiently comparable to conduct robust DEA, data should be measured without errors, all relevant exogenous differences should be accounted for, and the remaining heterogeneity in the sample should be addressed through an effective outlier procedure. Therefore, the estimation of an efficient frontier is greatly influenced by the existence of outliers, especially in deterministic methods of benchmarking such as DEA.

In the context of TCB18, where the TSOs being assessed operate in very different operating environments and data errors are likely to occur (see section 3.1), outliers are likely to be present in the sample.

5.3.2 Sumicsid's approach

Sumicsid follows the same outlier procedure as the Bundesnetzagentur in the second regulatory period.[174] This approach was developed to fall in line with the legal requirements of the ARegV. In addition to the formal outlier tests outlined in the ARegV, Sumicsid removes one TSO from the sample ex ante as it states this TSO is 'almost always [...] an extreme outlier' in the various model specifications it tested.[175]

Other outlier tests are discussed in the main report,[176] but it is not clear if or how these have been used in the final analysis.

5.3.3 Critique and proposed solution

Sumicsid does not provide any rationale for its decision to use the outlier procedure specified in the ARegV. The purpose of the ARegV is to enable national comparisons among German regulated entities. It is neither legally binding nor sufficient in an international benchmarking context.

Indeed, the Bundesnetzagentur's outlier procedure has been challenged in the past as it is inconsistent with the academic literature and is biased against identifying outliers. Furthermore, Sumicsid mentions alternative approaches to outlier detection in the main report (such as an examination of DEA weights

[173] GTS is estimated to be 73.2% efficient under NDRS by Sumicsid. The scale efficiency, defined as $\frac{DEA_{CRS}}{DEA_{VRS}}$ is also estimated to be 73.2%. Given that GTS is the second largest TSO in the sample, the NDRS and CRS efficiencies are likely the same. If this is the case then the efficiency under VRS would be 100%. Note that this observation assumes that there is no interaction between the returns to-scale-assumption and the outliers detected under the different assumptions.
[174] The procedure in the third regulatory period was adapted slightly to account for the functional form assumption in the SFA.
[175] Sumicsid (2019), 'Pan-European cost-efficiency benchmark for gas transmission system operators main report', July, p. 36.
[176] Sumicsid (2019), 'Pan-European cost-efficiency benchmark for gas transmission system operators main report', July, p. 30.

and the econometric method)[177] but does not explain how or if these methods were used at any stage in the TCB18 study.

Specific issues relating to Sumicsid's dominance and super-efficiency tests are outlined below.

Dominance test

In an expert opinion on Sumicsid's dominance test in the benchmarking of German gas TSOs using DEA, we concluded that the dominance test is not based on any theoretical foundation and not legally consistent with the ARegV.[178] Legal consistency is not an issue in the TCB18 study as there are no binding legal requirements for this international benchmarking exercise. More importantly, however, the dominance test is not supported in the academic literature and suffers from a number of flaws.

- The dominance test requires the TSO efficiencies to follow a half-normal distribution, which is inconsistent with the non-parametric nature of the DEA method. Moreover, Sumicsid has not presented any evidence that the efficiency scores do indeed follow this distribution.
- The academic reference for the dominance test notes that his tests are asymptotically valid under the maintained assumptions, which means that they are appropriate only for large samples.[179, 180]
- Efficiencies of the same units are being compared in the numerator and denominator of the test statistic. This invalidates the comparison to the dominance test in the academic literature, which require independent samples of efficiencies to be compared.
- The F-distribution is an asymmetric distribution defined on a real line with positive values ranging from zero to infinity, and with no fixed upper bound. However the test statistic of the dominance test is bound between zero and one as a TSO cannot be more than 100% efficient or less than 0% efficient in a DEA model.

In summary, the dominance test has no theoretical basis and cannot be used to identify dominant units, regardless of the sample size.

Sumicsid, in its response to our expert opinion has, without exception, agreed with the four limitations highlighted in our expert report and summarised above.[181] Sumicsid specifically notes that:

> In summary, we consider the four objections towards the F-distribution raised by the Oxera note as valid.[182]

Nevertheless, Sumicsid (2019) argues that the test 'makes the most use of available information' and that it 'is a cautious test'.

[177] Sumicsid (2019), 'Pan-European cost-efficiency benchmark for gas transmission system operators main report', July, p. 31.
[178] Kumbhakar, S., Parthasarathy, S. and Thanassoulis, E. (2018), 'Validity of Bundesnetzagentur's dominance test for outlier analysis under Data Envelopment Analysis', August.
[179] Banker, R.D. (1993), 'Maximum Likelihood, Consistency and Data Envelopment Analysis: A Statistical Foundation', *Management Science* **39**:10, pp. 1265–73, Banker, R.D. (1996), 'Hypothesis tests using data envelopment analysis', *The Journal of Productivity Analysis*, **7**, pp. 139–59.
[180] Sumicsid uses sample size as a justification for ignoring the results from SFA (or, indeed, not estimating SFA models at all) yet it deems the sample size to be large enough to conduct this outlier test. In neither case does Sumicsid provide empirical justifications for its choices.
[181] Sumicsid (2019), 'Outliers in DEA based regulatory benchmarking response to the Oxera report', October.
[182] Sumicsid (2019), 'Outliers in DEA based regulatory benchmarking response to the Oxera report', October, p. 16.

In response to Sumicsid, we showed not only that the test lacks a theoretical foundation, but also that it is biased towards non-rejection (i.e. non-identification of potential outliers).[183] The claim made in Sumicsid's response to our initial report, that its test is cautious, does not stand up to scrutiny which we also evidenced empirically in our rejoinder.

For example, on the German gas TSO dataset (which also forms the majority of TCB18 Gas data), one of the gas TSOs acts as a peer for 12 out of 16 units. Excluding this TSO has a material effect on the industry's efficiency (7.8 percentage points) and on costs (€136.4m p.a.).[184] Moreover, the affected TSOs see their cost allowance reduced by 10.6 percentage points (€202.8m p.a.) if the unit is left in the dataset (as is the case with the application of Sumicsid's dominance test).[185] The particular TSO's maximum impact on an affected TSO's efficiency is as high as 22.3 percentage points. Nevertheless, such an influential unit is not identified as an outlier under the dominance test. This illustrates the shortcomings outlined above and demonstrates that the test is not cautious.

There are better alternatives for the dominance test, but they require further theoretical development. The most promising candidate is the bootstrap test proposed in our initial expert report. The bootstrap-based test is a non-parametric test consistent with the non-parametric nature of DEA, requires fewer assumptions than other non-parametric options available, and explicitly takes into account the pairing structure of efficiencies both with and without a potential outlier.[186]

Super-efficiency test

Although the super-efficiency test as applied by Sumicsid (unlike the dominance test) has a theoretical foundation, it is insufficient to detect all abnormally super-efficient units. In particular two issues are of importance in the current case.

- **Outliers are unlikely to be identified in volatile samples.** The critical efficiency value, above which a TSO is identified as an outlier, is directly proportional to the inter-quartile range (IQR) of the estimated efficiency scores. Therefore, if the results from a DEA model are highly volatile (i.e. have a large range in estimated efficiency scores), this could increase the IQR on average, which would imply that the critical efficiency value would be higher for a unit to be identified as an outlier. In the present case for instance, due to the high volatility in the sample, a TSO would have to exhibit 40% super efficiency to be considered an outlier (i.e. it could increase its costs by 40% and still be considered efficient). Super efficiency values well below this critical value should be considered unusual and may warrant removal from the sample. For instance, the critical value for the super efficiency in German regulatory benchmarking (calculated with the same method) was in a range between 121.0% and 106.9% in the latest

[183] Kumbhakar, S., Parthasarathy, S. and Thanassoulis, E. (2019), 'Rejoinder to Sumicsid's response to the Oxera report on Bundesnetzagentur's dominance test in DEA', May.
[184] The average percentage impact is the mean of the difference in efficiencies for the remaining 15 TSOs after the removal of the dominance TSO. The total monetary impact based on the average percentage impact is the average percentage impact × total TOTEX in the industry.
[185] The average percentage impact for the affected TSOs is the mean of the difference in efficiencies for the TSOs that saw an increase in efficiency following the removal of the dominant TSO. The monetary impact on the affected TSOs is the sum of the monetary impacts for the each affected TSO: for TSO i, $\left(Eff(K\backslash i) - Eff(K)\right) \times TOTEX$.
[186] Kumbhakar, S., Parthasarathy, S. and Thanassoulis, E. (2018), 'Validity of Bundesnetzagentur's dominance test for outlier analysis under Data Envelopment Analysis', August.

regulatory period.[187] A high-volatility sample (and thus a high threshold for outliers) could be a result of unidentified heterogeneity, data errors or an imperfect model.

- **Sumicsid's analysis is vulnerable to masked outliers.** For example, there could be two network operators that are similar in characteristics and far removed from the rest of the sample, but one is masked (covered or hidden) by another. This is referred to as a 'masked' outlier. In such a case, a mechanistic application of the super-efficiency test could fail to identify either one as an outlier. Note that the same issue can also occur with respect to the dominance test.

Only a sequential application of the super-efficiency outlier tests that allow for the possibility of masking could reveal two or more 'joint' outliers. We suggest that a sequential exploration of outliers should be pursued. Such a sequential exploration has precedent in academic[188] and regulatory[189] literature. In its basic form, this involves:

- setting an 'absolute' threshold for super-efficiency, which can be based on operational expectations or regulatory expert view or values from alternative approaches;
- units with a super-efficiency value exceeding that threshold are removed;
- super-efficiencies are computed again on the reduced sample and newly identified super-efficient units that exceed the threshold are also removed;
- this sequence is followed until no new super-efficient outliers are identified (i.e. an iterative application).

After each step, conspicuous companies based on the absolute super-efficiency threshold are excluded and the remaining companies are investigated for additional abnormalities. Moreover, to allow for the possibility of 'joint' outliers, companies initially found not to be outliers should be tested again after removing their peers (i.e. efficient companies they are referenced against) temporarily for the purposes of the test. The process can be stopped when no company has a conspicuously high super-efficiency value.

A sequential application of super-efficiency outlier tests would result in the identification of more outliers, resulting in a more homogeneous sample for efficiency estimation.

Once an outlier TSO has been identified, its data should be re-examined to ensure that it is reported consistently with the other TSOs and without error. The model specification should also be re-visited to ensure that it is not biased in favour of the outlier TSO, or against the non-peer TSOs. We understand that Sumicsid did not scrutinise the data or model specification upon detecting outliers in its model.

It is important to note that all outlier procedures rely on assumptions and are contingent on the model specification. Furthermore, most tests require

[187] Frontier Economics (2019),' Effizienzvergleich Verteilnetzbetreiber Gas (3. RP) Gutachten für die Bundesnetzagentur', May p. 129, Sumicsid and Swiss Economics (2019),'Effizienzvergleich Verteilernetzbetreiber Strom der dritten Regulierungsperiode (EVS3) Gutachten', April, p. 103.
[188] Thanassoulis, E, (1999) 'Setting Achievement Targets for School Children', *Education Economics*, **7**:2, pp. 101–19.
[189] Deuchert, E. and Parthasarathy, S. (2018–19), five-part series of articles on the German energy regulator's benchmarking framework covering efficiency methods (DEA and SFA), functional form assumptions, cost driver analysis, outlier analysis and model validation, ew–Magazin für die Energiewirtschaft.

somewhat arbitrary thresholds to determine whether a particular observation is an outlier. For this reason, it is essential that the data collection and construction process is robust to identify data errors. Moreover, the final model used to estimate efficiency scores must appropriately capture differences in operation characteristics across TSOs. In other words, even a robust and academically valid outlier procedure is not a replacement for proper data processing and model development.

5.4 Second-stage analysis

5.4.1 Description of the issue

If relevant cost drivers are omitted from the model specification, the resulting efficiency scores will be biased for TSOs which are especially affected by this cost driver. The relevance of potential cost drivers is commonly assessed with the cost driver analysis in the model-development phase of a benchmarking study and validated further through extensive sensitivity analysis.

5.4.2 Sumicsid's approach

Sumicsid uses second-stage regressions to test for the exclusion of relevant cost drivers. It is not clear from the individual report whether the regression is estimated using OLS, ROLS or some other estimator, but Sumicsid states in the main report that:

> second-stage analyses are typically done using graphical inspection, non-parametric Kruskal-Wallis tests for ordinal differences and truncated Tobit regressions for cardinal variables.[190]

Sumicsid further states that such second-stage analysis of this sort is 'routinely done' to identify omitted cost drivers.[191] This analysis is presented in the individual report.

5.4.3 Critique and proposed solution

The second-stage analysis forms the core of Sumicsid's model validation in the individual reports.[192] However we are not aware of any literature that specifically focuses on using DEA, followed by regression, to justify a set of input–output variables.[193] Second-stage regressions are sometimes used to adjust efficiencies from a first stage DEA model for environmental factors;[194] however, this serves a different purpose to that stated by Sumicsid. Furthermore, the second-stage analysis used in the literature is valid if the following conditions hold.

- The location of the efficient frontier based on the first-stage analysis is **not** affected by the variables used in the second-stage analysis.[195]

[190] Sumicsid (2019), 'Pan-European cost-efficiency benchmark for gas transmission system operators main report', July, para. 4.09.
[191] Sumicsid (2019), 'Project TCB18 Individual Benchmarking Report GTS – 209', July, p. 35.
[192] For example, see Sumicsid (2019), 'Project TCB18 Individual Benchmarking Report GTS – 209', July, chapter 5.
[193] The Bundesnetzagentur has used such analysis in the past to identify omitted cost drivers. but such use is not supported by academic literature and was strongly challenged by the industry.
[194] For example, see the Norwegian Water Resources and Energy Directorate, 'Guidelines for revenue cap calculation in R', section 4.2.
[195] See Simar, L. and Wilson, P.W. (2007), 'Estimation and inference in two-stage, semi-parametric models of production processes', *Journal of Econometrics*, **136**:1, pp. 31–64, Simar, L. and Wilson, P.W. (2011), 'Two-stage DEA: caveat emptor', *Journal of Productivity Analysis*, **36**:205. Banker and Natarajan (2008) adopt a different variant of separability, as noted below. See Banker, R.D. and Natarajan, R. (2008), 'Evaluating contextual variables affecting productivity using data envelopment analysis', *Operations Research*, **56**:1, pp. 48–58.

- The first- and second-stage variables should be independent of each other, although they can be correlated within first- or second-stage variables.
- Estimates of efficiency from the first stage should first be adjusted for the serial correlation bias, because the dataset in the first stage is finite. The bias correction can be estimated using a bootstrapping procedure.[196]

These conditions do not hold in TCB18. The vast majority of variables tested in the second stage analysis are scaled measures of NormGrid and pipe length which are highly dependent on the first-stage variable NormGrid weighted with slope factors and pipe length scaled with humidity factors. Therefore, they are not independent as required. Only two variables tested (near_coast and dist_coast) might meet the requirements set out above. Even for those variables, the stated purpose of the second-stage analysis is different from the purpose for which it was developed and it is not clear if a correction for serial correlation bias was attempted.

As the first-stage DEA model does not offer information on statistical significance of the outputs and Sumicsid's second-stage analysis for further model validation is erroneous, there is a need to provide additional robustness checks to see what happens to the DEA efficiency scores if some of the omitted variables are used alongside others in the first stage. This could identify TSOs that are particularly susceptible to certain cost drivers, even if those cost drivers are not significant at the industry level. For example, densely populated area may be a cost driver which is specific to GTS and not a concern for many other TSOs operating in less densely populated areas.

In our review of TCB18 electricity, we used an empirical example to illustrate the invalidity of Sumicsid's approach.[197] We first estimated the efficiencies of TSOs with two of Sumicsid's cost drivers, excluding one. We then tested whether the omitted cost driver would be identified as such using Sumicsid's second-stage analysis. As Sumicsid makes a number of claims with respect to its final model, including the importance of each cost driver in explaining costs,[198] one would expect all cost drivers to be statistically significant at the second stage—especially adjusted NormGrid, given that Sumicsid states that this is the main cost driver. However, for two out of three cost drivers in TCB18 electricity, the omitted cost driver was statistically insignificant and would not be identified as an omitted variable in Sumicsid's second-stage analysis.[199]

Similar inadequacies can also be demonstrated in other regulatory settings. For example, when applying Sumicsid's second-stage analysis and the method outlined above in German gas DSO benchmarking, none of the cost drivers used would have been identified as omitted if the model was estimated without them. In German electricity DSO benchmarking, only one cost driver would have been identified as significant and of the right sign.

These results illustrate that the second-stage validation used by Sumicsid is not able to identify relevant omitted cost drivers, even under favourable conditions. It cannot be argued, therefore, on the basis of this procedure, that

[196] See Simar, L. and Wilson, P.W. (2007), 'Estimation and inference in two-stage, semi-parametric models of production processes', *Journal of Econometrics*, **136**:1, pp. 31–64, Simar, L. and Wilson, P.W. (2011), 'Two-stage DEA: caveat emptor', *Journal of Productivity Analysis*, **36**:205.
[197] Oxera (2020), 'A critical assessment of TCB18 electricity', April, section 5.4.3.
[198] Sumicsid (2019), 'Pan-European cost-efficiency benchmark for electricity transmission system operators main report', July, pp. 32–34.
[199] Oxera (2020), 'A critical assessment of TCB18 electricity', April, Table 5.2.

no relevant cost drivers were omitted from the gas model, nor that the first-stage cost model is validated.

5.5 Dynamic efficiency and frontier shift

5.5.1 Description of the issue

Dynamic efficiency relates to how the efficiency and productivity of TSOs changes over time, while frontier shift relates to the ability of the most efficient operators in an industry to improve productivity. In a DEA context, frontier shift can be estimated by assessing the evolution of the efficient frontier over time. Alternatively, in an SFA context, frontier shift can be estimated by including time variables (e.g. a time trend or time dummies) in the model specification. The assessment of frontier shift is a critical aspect of regulatory benchmarking, as the frontier shift productivity improvements (or deteriorations) can be achieved by all companies in an industry.

Furthermore, the estimated frontier shift in the sample could be used to validate the model. For example, if the estimated frontier shift is significantly higher or lower than what is expected, it could point to flaws in the cost or cost driver construction, or the inability of the model to account for expenditure cycles.

5.5.2 Sumicsid's approach

Sumicsid does not discuss frontier shift in its main report or associated appendices. However, it has published the frontier shift results in a subsequent report.[200]

Sumicsid uses a DEA Malmquist approach to estimate the rate of technical efficiency improvement (i.e. 'catch-up'), frontier shift and general productivity improvements in the sample of TSOs.[201] While the overall model is the same as that used to estimate static efficiency in the main report, the German TSOs' data is excluded as data was only available for one year (and therefore a rate of change cannot be calculated).

The analysis indicates that there has been no productivity growth in the sample (on average) and the frontier has regressed at a rate of 1.7% p.a. That is, efficient real costs are increasing at a rate of 1.7% p.a. However, Sumicsid heavily caveats the conclusions of its analysis. For example, it states:

> The results with respect to the best-practice frontier should not be overinterpreted given the relatively low number of peers in the dynamic study.[202]

5.5.3 Critique and proposed solution

There are several key aspects to frontier shift analysis that Sumicsid has not discussed in its dynamic efficiency report that may limit the robustness of Sumicsid's frontier shift estimate, although we do not discuss these in detail in

[200] Sumicsid (2020), 'Project CEER-TCB18 Dynamic efficiency and productivity changes for gas transmission system operators MAIN REPORT', April.
[201] The application of a DEA-based Malmquist method to measure indices of TFP change—decomposable further into technical efficiency change, scale efficiency change and technological change—follows the methods developed in Färe, R., Grosskopf, S., Norris, M. and Zhang, Z. (1994), 'Productivity Growth, Technical Progress, and Efficiency Change in Industrialized Countries', *American Economic Review*, **84**, pp. 66–83, and Ray, S.C. and Desli, E. (1997), 'Productivity Growth, Technical Progress, and Efficiency Change in Industrialized Countries: Comment', *American Economic Review*, **87**, pp. 1033–39.
[202] Sumicsid (2020), 'Project CEER-TCB18 Dynamic efficiency and productivity changes for gas transmission system operators MAIN REPORT', April, p. 20.

this report.[203] Instead, we focus on a number of inconsistencies in Sumicsid's interpretation of the analysis, and why the frontier shift analysis casts doubt on the robustness of the TCB18 model for the static analysis.

For example, unlike in the static efficiency analysis, Sumicsid caveats its dynamic efficiency calculations in the report.

> In the particular case for gas transmission, the frontier shift results are moreover affected by the low number of peers in each year (3-4), making the frontier generalizations less robust than the efficiency change estimates analysis, stating that the small dataset makes the conclusions less robust. [204]

It is true that, in general, a larger sample will lead to more robust estimates of any parameter. However, Sumicsid does not acknowledge that:

- there are only four peers in the static efficiency comparisons—if the frontier shift estimate is not reliable because there are only 3–4 peers, then it's not clear why the static efficiency comparisons should be any more reliable;
- the reliability of the estimated frontier impacts both the estimate of frontier shift and the estimates of catch-up (efficiency change). Indeed, we see that the only catch-up observed in the sample is driven by the estimated frontier regressing to the performance of the 'inefficient' TSOs, rather than the 'inefficient' TSOs improving their cost performance;
- the dataset Sumicsid uses in the dynamic efficiency analysis is significantly larger than the dataset used in the static analysis—Sumicsid only uses data from 2017 (or 2015 for the German TSOs) in the static analysis, which limits the sample size to 29 observations. Meanwhile, in the dynamic analysis, the data consists of 9–11 TSOs[205] over 5 years, which equates to 45–55 observations.

Sumicsid's estimated frontier regress of 1.7% p.a. is large in magnitude and of the opposite sign to what is commonly estimated (and expected) in regulatory applications. Sumicsid attempts to caveat the unexpected estimate by stating that the frontier shift is estimated relative to general inflation and therefore represents the performance of the TSO sector relative to the wider economy.[206] However, frontier shift targets are typically estimated relative to CPI (or some other general price inflation), so there is nothing unique about the TCB18 study in this respect and this cannot be considered a sufficient explanation for the unintuitive results.

An alternative and more realistic explanation for such a large and negative estimated frontier shift is that the TCB18 model is mis-specified. In particular, if Sumicsid's cost drivers do not explain changes in expenditure over time, relevant cost drivers that may help to explain changes in costs over time may

[203] For example, it is not clear how or if Sumicsid adapted its static outlier approach (which is itself fundamentally flawed even in a static context) to account for 'dynamic' outliers. Moreover, the appropriate returns-to-scale assumption is not discussed anywhere in the report. This is concerning as a CRS assumption must be used when estimating frontier shift, regardless of the 'true' returns to scale in a static context (see Thanassoulis, E. (2001), *Introduction to the Theory and Application of Data Envelopment Analysis: A foundation text with integrated software*, Kluwer Academic Publishers, pp. 177–178, and Färe, R., Grosskopf, S. and Margaritis, D. (2008), 'Efficiency and productivity: Malmquist and more', *The measurement of productive efficiency and productivity growth*, **5**, pp. 522–622). As with the static analysis, the dynamic efficiency analysis has not been validated against other estimation approaches such as SFA.
[204] Sumicsid (2020), 'Project CEER-TCB18 Dynamic efficiency and productivity changes for gas transmission system operators MAIN REPORT', April, Executive Summary.
[205] Sumicsid (2020), 'Project CEER-TCB18 Dynamic efficiency and productivity changes for gas transmission system operators MAIN REPORT', April, p. 16.
[206] Sumicsid (2020), 'Project CEER-TCB18 Dynamic efficiency and productivity changes for gas transmission system operators MAIN REPORT', April, p. 20.

be missing, and the position of a TSO in the investment cycle may also impact the estimated efficiency scores.

The inability of the DEA model to account for changes in efficient expenditure over time also raises additional concerns regarding the ability of the model to account for differences in efficient expenditure across TSOs. For example, if the model is unable to capture the general trend of increasing regulatory burden over time, then it is unlikely that the model can capture differences in regulatory burden between TSOs. In this case, a TSO's estimated efficiency score will be driven by differences in regulatory burden as well as (or rather than) genuine differences in efficiency.

Importantly, it appears to be inconsistent to assess one result from the TCB18 study (the static analysis) to be more robust than another (the dynamic analysis). Both are intrinsically related, and both are estimated using the same model and similar datasets. Sumicsid (and the project participants) should conclude that either (i) both the static and dynamic analyses are of high quality and can be used in regulatory applications; or (ii) both the static and dynamic analyses are not robust and cannot be used. This report has demonstrated that the latter conclusion is most appropriate in this instance.

6 Conclusion

International benchmarking can be a powerful tool for companies and regulators to assess the efficiency of TSOs. This is especially true in the context of the gas transmission industry, where the sector is often characterised by national monopolies, thus making national benchmarking (outside of Germany, where multiple TSOs operate) challenging. In this sense, we welcome projects such as the TCB18 study and e2gas, which have attempted to develop a framework for the regular assessment of TSOs.

Nevertheless, the TCB18 study itself suffers from a number of weaknesses at every stage of the benchmarking process that mean the estimated efficiency scores cannot be interpreted as 'true' differences in efficiency in their current form. Specifically, the estimated efficiency scores from TCB18 **cannot be used for regulatory purposes** (or operational or valuation purposes).

We have identified several areas of the benchmarking project that require significant work before the results can be fit for purpose.

- **Transparency**. Sumicsid's level of transparency falls short of what is commonly considered good-practice in regulatory contexts. The statements it makes are often not substantiated with evidence. In addition, the outputs do not contain relevant information for TSOs to be able to improve their efficiency.
- **Data collection and construction**. The process that generated the dataset on which the final model was derived is not sufficient to ensure that all data is reported consistently across TSOs. This issue is exacerbated by the inclusion of German TSOs' data, as the German TSOs did not actively participate in the study. Given that multiple errors were detected in our review of TCB18 electricity, and the data collection process in TCB18 gas is very similar to that of TCB18 electricity, data errors are also likely to be present in the gas data. Furthermore, the adjustments for price-level differences and operating environments that Sumicsid has made to the data are insufficient to capture the heterogeneity across TSOs.
- **Model development**. Sumicsid's cost driver analysis is not well-documented, and the analysis that Sumicsid does present indicates that the model is unlikely to be robust. Indeed, the model clearly violates operational intuition and is arbitrarily restricted to a small number of cost drivers. Thus, relevant cost drivers (such as population density) are likely to have been omitted from the model, and the resulting efficiency scores are likely to be biased.
- **Application and validation**. The assumptions that fed into Sumicsid's application of DEA are not justified. Unlike in previous iterations of the study, no attempt has been made to validate results using different benchmarking methods, such as SFA. The outlier procedure is not consistent with academic practice and not justified for the current context. Furthermore, Sumicsid's approach to validating its model through second stage-analysis is neither theoretically valid nor sufficient to detect omitted cost drivers.

These issues cannot be corrected through simple refinements to the current model. We provide a number of recommendations in Appendix A2 of this report to improve the analysis for future iterations of the study. In the absence of significant refinements to the TCB18 study, one must conclude that the results from the analysis cannot be used in any setting.

By incorporating the recommendations presented in this report, the CEER and Sumicsid (or any future consultant) will be better able to develop a robust model (or set of models) for cost benchmarking. In this regard, it can also be helpful to consider debriefs on process and methodology, involving all parties, to help future studies.

A1 Summary of Sumicsid's TCB18 approach

This section gives an overview of the TCB18 study to provide context for the issues outlined in this report. Note that the approach taken by Sumicsid in TCB18 gas is very similar to the one taken in TCB18 electricity.[207]

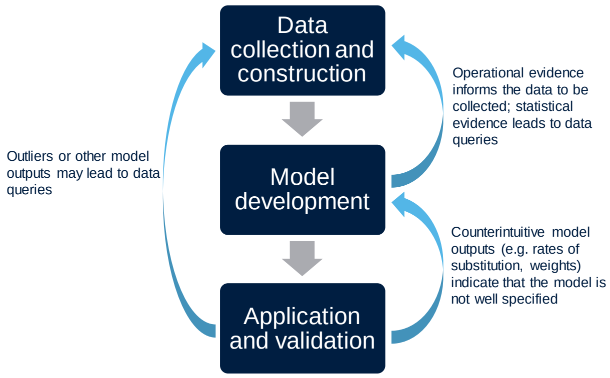
An efficiency benchmarking exercise can broadly be divided into three phases.

1. **Data collection and construction.** Here, data from TSOs, and in the case of the German TSOs the regulator, is collected, audited and screened for errors (such as misreporting or anomalies). After one is confident that the data is (to a reasonable degree) without errors, it needs to be normalised for differences in reporting (e.g. accounting guidelines can vary, expenditure needs to be converted into a single currency), regulatory frameworks and operational characteristics to ensure that the cost base is comparable across TSOs.
2. **Model development.** Given the data, the model for benchmarking is derived based on a combination of scientific procedure and expert judgement. This concerns the definition of costs (e.g. TOTEX), cost drivers (e.g. network length, environmental factors), the approach to the treatment of outliers given the chosen model, and the choice of benchmarking model, such as data envelopment analysis (DEA)[208] and stochastic frontier analysis (SFA),[209] and motivating the assumptions underpinning each step.
3. **Application and validation.** Once the model specification(s) and method are selected based on best available data, the model needs to be robustly estimated. This involves, among other things, validating the results from the model, as well as undertaking robust sensitivities to ensure that the results are not driven by specific assumptions made by the modeller.

In reality, this is not a sequential procedure but a highly iterative one, with feedback occurring between the various steps. Even the results from the final application of the benchmarking model may highlight additional data and modelling queries that necessitate further analysis. The process is illustrated in Figure A1.1 below.

[207] TCB18 gas also copies a considerable amount of text from e2gas, a previous version of the pan-European gas benchmarking study.
[208] DEA is a mathematical non-parametric approach that is widely used when benchmarking regulated companies. For a more detailed discussion on DEA, see Thanassoulis, E. (2001), *Introduction to the Theory and Application of Data Envelopment Analysis: A Foundation Text with Integrated Software*, Springer.
[209] SFA is an econometric approach to benchmarking regulated companies. For a more detailed discussion on SFA, see Kumbhakar, S.C, Wang, H-J and Horncastle, A. P. (2015), *A Practitioner's Guide to Stochastic Frontier Analysis Using STATA*, Cambridge University Press.

Figure A1.1 Benchmarking process



Source: Oxera.

We summarise each step taken by Sumicsid below.

A1.1 Sumicsid's approach to data collection and construction

In order to benchmark the cost performance of the European TSOs, it is essential to construct a homogenised dataset on the cost and outputs of the participating TSOs to enable a like-for-like comparison. That is, cost and outputs must be reported consistently (for example, in terms of allocating costs to specific activities), and the activities performed by the TSOs must be broadly similar.

Sumicsid states that it followed a six-stage approach to data collection and validation.[210]

1. **Asset system and audited financial statements.** This involved the collection of asset system data and audited financial statements of TSOs.
2. **Clear guides/templates.** CEER, Sumicsid and the TSOs worked interactively to establish reporting definitions to translate the data from the first stage into something that could be used for international benchmarking.
3. **Interaction (e.g. workshop).** There was interaction between the TSOs, NRAs and Sumicsid at all stages in the data-collection process to ensure the correct interpretation of reporting guidelines.
4. **National validation.** The national regulatory authorities (NRAs) validate the data to ensure the data is 'complete, consistent, correct and plausible'.
5. **Cross-validation.** Sumicsid performed an additional cross-validation exercise to identify and correct for any remaining misinterpretations of reporting guidelines.
6. **Data analysis.** Related to the cross-validation process, Sumicsid continued to review the data during the modelling process. The modelling process may detect errors or misreporting that were not identified in the previous steps.

[210] Sumicsid (2019), 'Pan-European cost-efficiency benchmark for gas transmission system operators main report', July, section 3.2.

As part of its final data checks, Sumicsid states that all TSOs participating in the TCB18 study received 'a dump of asset and financial files'[211] that they could review for missing or incorrect data. At this stage in the process, there were 'a few final corrections' for 'many' TSOs.[212]

The sixteen German gas TSOs did not participate in this process. Instead they participated 'through their NRA making previously validated data for the year 2015 available for the study'.[213] Sumicsid states that 'particular attention was paid to screen and analyze any potential differences between the reporting instructions in the previous benchmark, the national validation performed' for those TSOs' data.[214] Sumicsid does not explain the exact process they followed in this regard and the conclusions of its analysis (e.g. what differences were found, how were they corrected for).

For the TCB18 study, Sumicsid also collected data from external sources relating to inflation rates,[215] price-level differences[216] and environmental factors.[217]

Activities assessed in the benchmarking study

The management of a gas transmission network is an extremely complex operation. Sumicsid categorises each activity undertaken by TSOs as follows.[218]

- **Transport (T).** This entails the core service of TSOs in transporting energy from injection points to connection points (such as a downstream network, another TSO, or an end-client).
- **Grid maintenance (M).** Grid maintenance involves the repair of grid assets (both preventative and reactive) and the replacement of degraded equipment.
- **Grid planning (P).** This involves the analysis, planning and drafting of network expansion.
- **Indirect support (I).** Indirect expenditure includes administrative support functions (such as human resources and IT) that cannot be allocated to a specific activity. This includes central management costs.
- **System operations (S).** This involves assessing the real-time energy balancing, blending and injection of gases.
- **Market facilitation (X).** Market facilitation involves the facilitation or management of electricity marketplaces.
- **Offshore transport (TO).** This entails the transport of gas through offshore assets (e.g. undersea pipelines).
- **Gas storage (G).** This is the cost associated with maintaining gas storage facilities.

[211] Sumicsid (2019), 'Pan-European cost-efficiency benchmark for gas transmission system operators main report', July, p. 9.
[212] Ibid., p. 9.
[213] Ibid., p. 32.
[214] Ibid., p. 7.
[215] Ibid., Table 4.3.
[216] Ibid., Table 4.1.
[217] Ibid., Table 3.1.
[218] For more detail regarding how Sumicsid has defined these activities, see Sumicsid (2019), 'Pan-European cost-efficiency benchmark for gas transmission system operators main report', July, sections 4.4–4.12.

- **LNG Terminals (L).** This involves the operation and maintenance of LNG Terminals and peak shaving plants.
- **Other activities (O).** TSOs may perform activities that cannot be classified into the above nine activities.

Sumicsid considers T, M, P and a subset of I to be within the scope of the benchmarking study, and allocated costs and outputs on this basis. It claims that this was driven by 'analysis of common factors in cost reporting, the variability and homogeneity of the data and the separability of the activity'.[219]

Cost construction

Sumicsid assesses cost performance of the TSOs on a total expenditure (TOTEX) basis, where TOTEX is the sum of operating expenditure (OPEX) and capital expenditure (CAPEX). The OPEX is calculated on an annual basis in real terms (i.e. taking into account general price inflation) with no smoothing across years. The German TSOs only have data from 2015 available. Their OPEX is thus indexed to 2017 using a measure of inflation in overall goods. Cost items within OPEX that are deemed uncontrollable or out-of-scope[220] are excluded from the analysis. Sumicsid applies an adjustment to labour costs using the price level index (PLI) for civil engineering works to account for differences in wage rates across the TSOs. Finally, the OPEX for TSOs operating outside the euro area is converted to euros using the average exchange rate in 2017 (for all years).

CAPEX is typically 'lumpy'; a TSO may construct a large segment of the network in a particular year, and this will be registered as a large increase in investment. To avoid the efficiency estimates being driven by a TSO's position in the investment cycle, regulators often consider methods of 'smoothing' CAPEX.

In this case, Sumicsid has taken the annuity approach to CAPEX measurement, whereby the cost of the investment in a particular asset is spread over the asset's assumed lifetime. Specifically, Sumicsid uses investment stream data from 1973, assumed certain 'techno-economic lifetimes' of each asset, and assumed a real interest rate of 3% to calculate the annuity on each asset.[221] Assessed CAPEX in a particular year is the sum of the annuities for assets within the scope of benchmarked activities. It is not clear whether the German gas TSOs data was normalised using the same procedure.

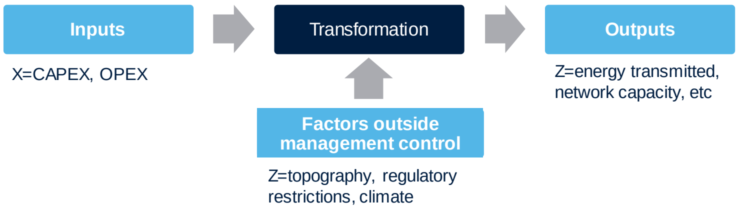
A1.2 Sumicsid's approach to model development

Cost benchmarking requires the comparison of homogenous units. However, TSOs operate at different scales, produce different mixes of outputs, and have different regulatory requirements and operating conditions (such as differences in topography and demography). In order to account for such severe heterogeneity, it is essential that the drivers of (efficient) expenditure are

[219] Sumicsid (2019), 'Pan-European cost-efficiency benchmark for gas transmission system operators main report', July, p. 18.
[220] These include energy costs, landowner compensation, right-of-way and easement fees, taxes and levies, depreciation, research and development and the rent of the main office building
[221] $A = CAPEX * \frac{r}{(1-(1+r)^{-T}}$, where A is the annuity, r is the assumed interest rate and T is the assumed lifetime of the asset. CAPEX is the expenditure for the asset. This formula splits CAPEX into T constant payments of A. For example an investment of 1,000,000€ in an asset with a lifetime of 50 years at an interest rate of 3% would result in an annuity of $A = 1{,}000{,}000€ * \frac{0.03}{1-(1+0.03)^{-50}}$ =38,865.49€ per year for 50 years.

sufficiently understood and controlled for. In this context, we assess the process shown in Figure A1.2 below.

Figure A1.2 Production process



Source: Oxera.

Relevant cost drivers (outputs or environmental factors) are included in the benchmarking model in order to homogenise the characteristics of different TSOs. Sumicsid states that in an ideal setting, the cost drivers should be:

- exogenous—i.e. outside of management control;
- complete—i.e. accounting for all operating characteristics;
- operable—i.e. clearly defined and measurable;
- non-redundant—i.e. the set should be as small as possible to avoid unnecessary duplication.[222]

While Sumicsid does not discuss this in the published report, it notes in some workshop slides that it has used econometric analysis to test the relevance of candidate cost drivers. In particular, Sumicsid states it has used 'robust OLS' (ROLS) to estimate and validate the relationship between costs and cost drivers in the main report.[223] ROLS is an extension of the OLS estimator, which attaches less weight to observations that are further from the regression line[224] (i.e. observations that fit the assumed model less well). Based on Sumicsid's final report, it appears to have particularly focused on the model-fit of such cost drivers (i.e. the extent to which variations in the cost drivers in the model can explain variations in costs) when determining the appropriate outputs to use in its benchmarking model.[225]

The final outputs used in Sumicsid's benchmarking model are the following.

- Normalised Grid ('NormGrid') adjusted for slope factors—a measure of the assets used to deliver outputs, adjusted for the topography (i.e. slope)

[222] See Sumicsid (2019), 'Pan-European cost-efficiency benchmark for gas transmission system operators main report', July, p. 29.
[223] Ibid., Table 5.4.
[224] Regression analysis is a statistical method that separates out the impact of different factors in explaining movements in the key variable of interest. This variable of interest is known as the 'dependent variable' (TOTEX in this context). The regression identifies how the different cost drivers contribute to explaining the observed values of the dependent variable, and whether the cost drivers are statistically significant. Simple regression, with one dependent variable and one cost driver, can be visualised graphically as a 'line of best fit'. This visualisation becomes more difficult as more cost drivers are added to the model.
[225] For example, Sumicsid presents only the adjusted R^2 for three models in a table in its report (see Sumicsid (2019), 'Pan-European cost-efficiency benchmark for gas transmission system operators main report', July, Table 5-4). In addition, the workshop slides tended to focus on the adjusted R^2 for the validation of NormGrid (see Sumicsid (2018), 'Validation of NormGrid and Preliminary Environmental Results', November, slides 6–11).

characteristics in the TSO's service area. This output is discussed in more detail below.

- Sum of connections—the sum of the connections across all types of connection points owned by the TSOs.
- Compressor power—the sum of installed power in compressors, irrespective of type.
- Weighted pipes—total length of pipelines multiplied with the humidity environmental factors.

NormGrid construction

NormGrid is an aggregated measure of the in-scope assets deployed by TSOs. Sumicsid describes it as 'a cost-norm for the construction costs for the standard assets'.[226] Specifically, NormGrid is calculated as the weighted sum of in-scope assets according to the equations below.

$$NormGrid = NormGrid_{OPEX} + NormGrid_{CAPEX}$$

$$NormGrid_{OPEX} = \sum_{t}\sum_{a}(N_{at} * w_a)$$

$$NormGrid_{CAPEX} = \sum_{t}\sum_{a} n_{at} * v_a * \alpha(r, T_a)$$

Where:

- N_{at} is the number of assets of type a, acquired at time t;
- w_a is the OPEX weight for assets of type a;
- n_{at} is the number of assets of type a, acquired at time t and in prime age;
- v_a is the CAPEX weight for assets of type a;
- α is the annuity function, which is determined by:
  - the real interest rate, r;
  - the 'techno-economic life' of the asset, T_a.

The weight on each asset is supposed to account for the characteristics of the asset within an asset class. For example, it is typically more costly to construct and operate pipeline with a larger diameter than a pipeline with a lower diameter. Sumicsid outlines the formula for each asset weight in an appendix to its main report.[227] It is not clear whether the NormGrid variable for the German gas TSOs was constructed in the same manner, as detailed asset data is not usually part of German TSO benchmarking.

Environmental adjustment

Sumicsid states that 'environmental factors, mainly related to the external environment of the pipeline' affect the construction costs of the pipelines.[228]

[226] Sumicsid (2019), 'Pan-European cost-efficiency benchmark for gas transmission system operators main report', July, p. 25.
[227] Sumicsid (2019), 'Norm Grid Development', July, section 2.
[228] Sumicsid (2019), 'Norm Grid Development', July, section 3.

Sumicsid considered factors that could account for these differences in operating characteristics, including land use type, topography, vegetation type, soil humidity, subsurface features, extreme temperatures, and salinity. In its final report, Sumicsid states that 'extensive statistical tests revealed correlations and interaction between several of the factors', and that 'the most important factor for gas was topography (slope class)'.[229]

Sumicsid adjusted the NormGrid variable to account for land use type by multiplying it by an overall environmental factor. This factor is calculated by multiplying the share of a TSO's service area covered in a certain feature with a complexity factor. We illustrate the calculation of this factor for GTS's operating area in Table A1.1 below.

Table A1.1 Calculation of the environmental adjustment (slope)

Feature	Factor (A)	Share of area covered (B)	Resulting adjustment (A*B)
Flat	1	97.0%	0.97
Undulating	1.15	2.6%	0.03
Hilly	1.35	0.4%	0.01
Mountain	2.25	0%	0
Sum		100%	1.01

Source: Oxera analysis, based on Sumicsid data.

Sumicsid also adjusts the pipe length variable with a different environmental factor based on the humidity complexity factor. The land use vegetation and subsoil factors developed remain unused.

A1.3 Sumicsid's approach to efficiency estimation and model validation

Sumicsid uses DEA[230] to estimate the relative efficiency of the TSOs. DEA uses linear programming models to estimate the minimum level of TOTEX needed for a TSO to meet its outputs (i.e. the cost drivers considered in the model) given the cost and output data for all the TSOs. The relatively efficient TSOs would be those for which no better performing TSO ('peer') could be identified in the dataset, and these TSOs would form the efficient frontier. DEA is further explained in Box A1.1.

[229] Sumicsid (2019), 'Pan-European cost-efficiency benchmark for gas transmission system operators main report', July, p. 34.
[230] Thanassoulis, E. (2001), *Introduction to the Theory and Application of Data Envelopment Analysis: A foundation text with integrated software*, Kluwer Academic Publishers.

Box A1.1 An overview of DEA

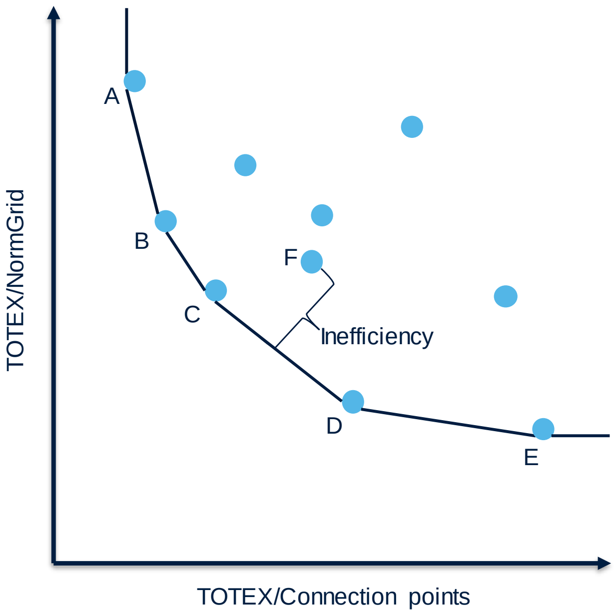
DEA has been used by a number of regulators and academics to assess the efficiency of companies and public service providers (such as in health and education). In its simplest form, DEA is an intuitive and transparent method for estimating the efficient frontier.

Suppose that all companies are identical in their production volume—for example, all firms serve the same number of customers who live in a comparable area. All one needs to do is order firms by their cost, and the one with the lowest cost sets the benchmark (it also has the lowest unit cost) or the efficient frontier.

However, the reality is far more complicated. The TSO dataset is characterised by severe heterogeneity, and a simple cost-by-cost comparison is not feasible. Multiple outputs are involved, and they differ across the TSOs. In this context, for an individual TSO, DEA tries to identify a peer TSO from existing (efficient) ones by combining them in certain proportions by means of linear programming. The peer TSO produces the same or more output than the TSO in question, but at lower cost (i.e. a lower level of TOTEX).

Figure A1.3 presents a stylised example of DEA in which NormGrid and connection points are the two outputs, TOTEX is the input, and the technology exhibits constant returns to scale. TSOs A, B, C and D are identified as efficient as no other TSO in the sample produces more of any one output without producing less of another output for a unit of TOTEX. TSO E is estimated to be inefficient, as TSO B can produce more of both outputs for the same unit of TOTEX. Similarly, TSO F is estimated to be inefficient, as TSOs B and C can produce more outputs for the same unit of TOTEX. DEA can be extended to account for any number of inputs or outputs.

Figure A1.3 Stylised example of DEA



Note: It is assumed that interpolation between real TSOs leads to a feasible point for any TSO to operate. Therefore each point on the graph enveloped by the frontier ABCDE and the frontier itself represents a real or virtual TSO. ABCDE is the efficient frontier in that no real or virtual TSO can exceed one of its outputs without attaining less on the other output for a unit of TOTEX.

Source: Oxera.

Returns to scale

One of the key steps in the application of DEA relates to the specification of the returns-to-scale assumption. This is explained in Box A1.2.

Sumicsid assumes non-decreasing returns to scale (NDRS)[231] in the model. NDRS can be seen as a variant of variable returns to scale (VRS), whereby TSOs may have increasing returns to scale when scale size is small (i.e. an increase (or decrease) in cost results in a more-than-proportionate increase (or decrease) in outputs), but operating under constant returns to scale (CRS) beyond a certain scale (in other words, no portion of the efficient frontier exhibits decreasing returns to scale). Sumicsid claims that this returns-to-scale assumption is supported by statistical evidence, but it does not present evidence to support this claim in its final outputs.

Box A1.2 Returns to scale

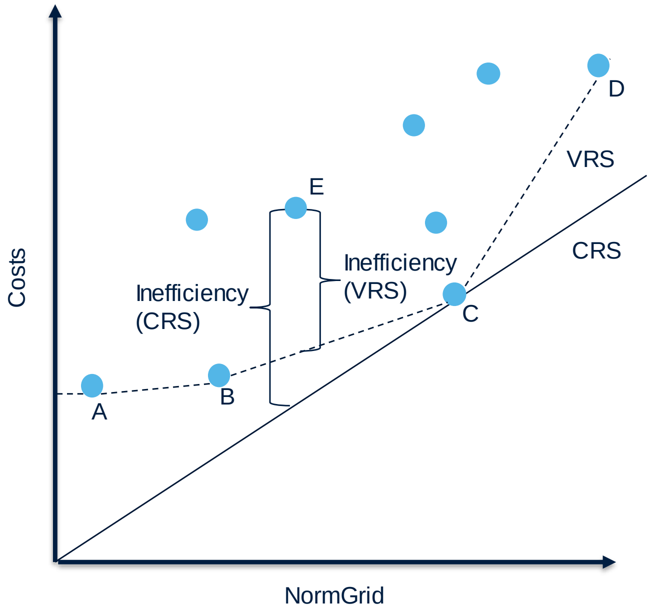
'Returns to scale' relates to how changes in outputs (e.g. NormGrid) are linked to associated changes in inputs (e.g. TOTEX) for efficient companies. In applications of DEA, the typical assumptions are constant returns to scale (CRS) or variable returns to scale (VRS).

Under CRS, it is assumed that when outputs rise or fall by a certain amount, say 5%, efficient companies would be expected to increase or decrease cost by that same amount (5%). In contrast, under VRS, when outputs rise or fall by a certain amount, say 5% again, efficient costs can rise or fall by a percentage greater than, less than, or equal to 5%, depending on whether decreasing, increasing or constant returns to scale are expected to prevail.

Figure A1.4 shows an illustrative example of different returns-to-scale assumptions in a single-input, single-output context. The line passing through C represents the CRS efficient frontier, while the line ABCD represents the VRS frontier. Under VRS, the line AB and BC exhibits increasing returns to scale in that for a unit rise in TOTEX (input) we have a more than proportional rise in NormGrid (output). In contrast, on CD we have decreasing returns to scale in that for a unit rise in TOTEX we have a less than proportional rise in NormGrid. The key practical implication is that C exhibiting maximum NormGrid per unit TOTEX cannot be scaled under VRS to provide benchmarks as it can along OB under CRS. VRS is therefore generally a less demanding assumption than CRS where benchmark performance is concerned.

[231] Otherwise known as 'increasing returns to scale' ('IRS').

Figure A1.4 Illustration of returns to scale



Source: Oxera.

Outlier analysis

The particular form of DEA that Sumicsid has considered in its analysis is deterministic. Such an approach takes no account of statistical noise (e.g. random data errors) in estimating the efficient frontier or individual efficiency scores. To mitigate the impact of potential outliers, Sumicsid follows the same outlier procedure as applied by the Bundesnetzagentur in the second regulatory period and outlined in the German Incentive Ordinance ('ARegV'). Specifically, Sumicsid performs a dominance and a super-efficiency test (explained in Box A1.3) to detect and remove outliers.

Box A1.3 The Bundesnetzagentur's outlier procedure

The Bundesnetzagentur is required to follow the methodology outlined in the ARegV to detect and remove outlier observations. In its application of DEA, the Bundesnetzagentur (i.e. or, specifically, the consultancy it has engaged) must remove the dominant and super-efficient outliers; these are defined below.

Dominance test

The aim of the dominance test is to identify companies that exert a substantial effect on the efficiencies of many other companies. The test, as outlined in the ARegV, compares the mean efficiency of all companies, including the potential outlier, with the mean efficiency calculated after excluding the potential outlier. If the efficiencies computed with and without the potential outlier are statistically different from each other at the 95% confidence level, the company is deemed dominant and removed from the sample.

By construction, removing one efficient company will increase the efficiencies of all companies to which it is a peer. To determine whether the difference in efficiencies with and without the company are statistically significant, the following test statistic is computed.

$$\frac{\sum_{k\in K\backslash i}(E(k;K\backslash i)-1)^2}{\sum_{k\in K\backslash i}(E(k;K)-1)^2}$$

Where:

- K is the total number of units (in this case, 16 TSOs);
- $E(k;K\backslash i)$ is the efficiency of TSO K estimated from the sample **excluding** the potential outlier, i;
- $E(k;K\backslash i)$ is the efficiency of TSO K estimated from the sample **including** the potential outlier, i.

This test statistic is assumed to follow an F distribution, and the company in question is removed if the value of the test statistic has a less than 5% chance of being randomly observed in the sample (i.e. a 95% confidence level).

Super-efficiency

The super-efficiency test aims to identify companies that are significantly more efficient than the rest of the sample. As defined in the ARegV, a company is considered super-efficient if its efficiency score when assessed relative to the rest of the companies (i.e. without it) exceeds the third-quartile efficiency value by more than 1.5 times the interquartile (i.e. between the third and first quartiles) range of efficiency values.

Furthermore, Sumicsid removed one TSO *before* the implementation of the outlier procedure. Sumicsid states that this TSO was 'almost always an extreme outlier' in its model-development phase,[232] but the empirical evidence for this was not provided in Sumicsid's final outputs.

Model validation

After estimating the model, Sumicsid performs a certain 'robustness analysis'[233] to test the sensitivity of the results to changes in the following

[232] Sumicsid (2019), 'Pan-European cost-efficiency benchmark for gas transmission system operators main report', July, p. 36.
[233] Sumicsid (2019), 'Pan-European cost-efficiency benchmark for gas transmission system operators main report', July, section 5.5.

specific modelling assumptions: interest rate, NormGrid calibration between OPEX and CAPEX, NormGrid weight for pipelines vs other assets, and salary corrections for capitalised labour in investments.[234] Sumicsid presents the impact of the changes on the average efficiency score in the sample. The impact on individual TSOs is not presented in the main report.

In the TSO-specific outputs, Sumicsid carries out second-stage analysis to test whether any relevant cost drivers have been omitted in the first-stage model used for efficiency estimation. Our understanding from the TSO-specific outputs is that in the second-stage analysis, Sumicsid regresses the estimated efficiency scores from the DEA model against potentially omitted drivers of expenditure (such as NormGrid weighted with land use factors) one at a time, according to the equation below.

$$Efficiency_i = \beta_0 + \beta_1 Y_i + \varepsilon_i$$

Where:

- $Efficiency_i$ is the estimated efficiency of TSO i;
- Y_i is the value of the omitted output of TSO i;
- ε_i is a random error component.

Sumicsid states that if the estimated coefficient $\hat{\beta}_1$ is statistically insignificant, the output in question, Y, is 'already considered in the model and do[es] not merit specific post-run corrections'.[235]

It is not clear from the individual report whether the regression is estimated using OLS, ROLS or some other estimator, but Sumicsid states in the main report that:

> […] second-stage analyses are typically done using graphical inspection, non-parametric Kruskal-Wallis tests for ordinal differences and truncated Tobit regressions for cardinal variables.[236]

Sumicsid further states that such second-stage analysis of this sort is 'routinely done' to identify omitted cost drivers,[237] but provides no evidence to support this statement.

In addition to the generic methods, one issue received further attention. To test whether different grid structures (i.e. meshed grid or point to point grid) had an impact on efficiencies Sumicsid tests a proxy, connection points per km pipeline. Sumicsid states that this proxy had 'no explanatory value on the DEA-score' or on the 'unit cost components'.[238] This suggests that the proxy was tested in second-stage regression and as an explanatory factor for unit costs, but not tested in the DEA model. Results of this analysis are not presented.

[234] Sumicsid (2019), 'Pan-European cost-efficiency benchmark for gas transmission system operators main report', July, p. 39.
[235] Sumicsid (2019), 'Project TCB18 Individual Benchmarking Report Fingrid – 131', July, p. 35.
[236] Sumicsid (2019), 'Pan-European cost-efficiency benchmark for gas transmission system operators main report', July, para. 4.09.
[237] Sumicsid (2019), 'Project TCB18 Individual Benchmarking Report Fingrid – 131', July, p. 35.
[238] Sumicsid (2019), 'Pan-European cost-efficiency benchmark for gas transmission system operators main report', July, para. 5.31.

A2 Recommendations for future iterations

Some of these weaknesses of TCB18, such as consistency in reporting guidelines, are partly driven by the lack of maturity in the international benchmarking process. We expect that some of these issues could be resolved with time and in future iterations of the study, as TSOs and NRAs become more familiar with the process. However, Sumicsid's concluding remarks in its main report are concerning, as they are not consistent with the significant issues and areas for future work identified through our review. For example:

> Regulatory benchmarking has reached a certain maturity through this process and model development, signaling both procedural and numerical robustness. Drawing on the work, the definitions and data standards as well as the model, CEER can readily plan for a repeated regular benchmarking at a considerably lower cost in time and resources, to the benefit of all involved. Although the current model brings improvements in particular in environmental factors, the inflation and salary corrections and the NormGrid definitions, the relative symmetry with the earlier model from E2GAS can be seen as a confirmation of the type of parameters and approaches chosen, leading to stable and predictable results. In this manner, the **future work can be directed towards further refinement of the activity scope and the interpretation of the results, rather than on the model development**.
>
> [emphasis added]

Areas which require significant future work include the following.

- Provide a clear conceptual (and, where possible, empirical) justification for any assumptions that feed into each stage of the benchmarking process.
- Relatedly, provide detailed description in the outputs and publish modelling codes (which can be anonymised) to aid in transparency.
- Establish an iterative data-collection procedure that ensures data is reported correctly and consistently across TSOs and validate these.
- If data of non-participating entities (e.g. TSOs) is to be used, the process by which the data is made comparable should be clearly laid out and any inconsistencies should be mentioned, along with the steps taken to address them.
- The sensitivity of any results to the inclusion of such entities should be tested and presented.
- Use statistical analysis, such as Monte Carlo simulations, to evaluate the impact of any potential data errors. This could then be used to adjust the estimated efficiency scores for setting cost allowances. Alternative evidence, such as SFA modelling, should also inform the extent of the adjustment.
- Robustly capture the impact of all input price differences on expenditure to avoid conflating efficiency and this exogenous factor.
- Perform a scientifically valid model-development process that: (i) is based on realistic modelling assumptions; (ii) tests the significance of alternative model specification; (iii) tests the sensitivity of the analysis to small changes in the sample: and (iv) avoids the arbitrary restriction of cost drivers to asset-based outputs.

- Relatedly, the analysis should not be too sensitive to the year in which efficiency is assessed. If the estimated efficiency of TSOs fluctuates significantly from year to year, the causes of this must be explored.
- If asset-based outputs are used, these must be validated through comparisons to alternative measures that proxy similar characteristics.
- Provide statistical evidence to support the modelling assumptions. In particular, its returns-to-scale assumptions must be justified and all of the relevant evidence should be presented.
- Develop a robust outlier-detection procedure based on academic and scientific best practice. This need not include exact tests recommended in this study (i.e. the bootstrap based dominance test and the iterative super-efficiency test); however, any assumptions that feed into the outlier tests should be clearly explained and supported.
- Analyse the outputs of a DEA model, such as cost driver weights, peers and lambdas, to ensure they are consistent with operational intuition.
- Avoid relying on second-stage validation to detect omitted cost drivers. In a DEA context, the impact of omitted cost drivers should be assessed by testing the sensitivity of the results to the inclusion of alternative cost drivers.
- Cross-check the analysis with alternative benchmarking methods, such as SFA, to validate whether the estimated efficiency scores can be attributed to genuine differences in efficiency or data uncertainty.
- Validate the static efficiency analysis through dynamic efficiency analysis. Not only is this an essential parameter in setting cost allowances, but it can also help to identify flaws with the model that are not evident from cross-sectional analysis.
- Interpret the results with a reasonable level of conservatism, given the general difficulties with top-down benchmarking in an international context and the specific flaws within TCB18.

By incorporating the recommendations presented in this report, the CEER and Sumicsid (or any future consultant) will be better able to develop a robust model (or set of models) for cost benchmarking. In this regard, it can also be helpful to consider debriefs involving all the parties on process and methodology to help future studies.

www.oxera.com