Quantonomics

QUANTITATIVE ECONOMICS

Memorandum

Date:	19/9/2023
From:	Michael Cunningham, Alice Giovani, Joe Hirschberg (Quantonomics)
To:	Adam Rapoport, Claire Preston, David Monk, Sasha Jergic, Su Wu, Dijana Cremona (AER/ACCC)
Subject:	Evoenergy – Benchmarking limitations

1 Introduction

In its Regulatory proposal for the ACT electricity distribution network 2024–29, Evoenergy makes a range of comments and criticisms relating to the benchmarking framework for distribution network service providers (DNSPs). These arguments are mostly directed to the econometric cost function modelling used in the Annual Benchmarking Reports and the AER's use of them for the purpose of assessing the efficiency of opex in the base year. The purpose of this report is to examine and evaluate Evoenergy's claimed shortcomings or limitations of the econometric analysis of opex efficiency and the use of the results in regulatory decision making.

1.1 Overview Evoenergy's positions

Sections 2 and 3 set out our understanding of Evoenergy's arguments on this topic. Here we briefly list the main arguments and where they are examined in this report.

(1) Evoenergy states that whilst the AER's benchmarking analysis is informative and important, it suffers from "many significant limitations" (Evoenergy, 2023a: 35). In light of its limitations, the AER should not view these efficiency estimates as highly precise or determinative. "Rather, the AER should consider its estimates of efficient opex as indicative at best" (Evoenergy, 2023a: 37). Related to these points, Evoenergy also notes that "The AER uses benchmarking and productivity analysis to measure the quantitative relationship between inputs used and outputs produced. To assess the efficiency of costs, base year revealed opex provides a reasonable estimate of the efficient and recurrent costs required to provide safe and reliable services while meeting regulatory obligations" (Evoenergy, 2023b: 20). Taken together, these points can be interpreted as suggesting that the AER should place less reliance on the results of benchmarking analysis when determining efficient base year opex. Since these

arguments are broader than those which follow, they are discussed at the end of the memo in section 4.

(2) The benchmarking models, Evoenergy argues, do not account for a sufficient range of operating environment factors (OEFs). There are "vast intrinsic differences in operating environments faced by different DNSPs" (Evoenergy, 2023a: 5). However, "the limited number of OEFs recognised by the AER and incorporated into the benchmarking analysis is inadequate to allow proper, like-with-like comparisons between DNSPs that would produce reliable estimates of efficient opex" (Evoenergy, 2023a: 6). OEFs are discussed in section 2 and this specific issue is discussed in section 2.3.

(3) Evoenergy also has reservations about the practice of making ex-post adjustments to efficiency scores derived from econometric modelling to account for the effects of specific OEFs. "The AER's extensive use of post-modelling OEF adjustments, rather than normalisation of costs for differences in DNSP operating environments before the benchmarking models are estimated, are likely to produce unreliable estimates of efficiency for individual DNSPs and potentially misidentify reference DNSPs" (Evoenergy, 2023a: 6). This issue is discussed in section 2.4.

(4) The opex cost function models continue to suffer from monotonicity violations and, Evoenergy states, Quantonomics has not found an approach to adequately address this (Evoenergy, 2023a: 6). Consequently, models with excess monotonicity violations are excluded when calculating the average efficiency score of a business. This affects comparisons of efficiency between DNSPs or over time, because average efficiency scores may be calculated using different sets of model results. "Hence, if a different set of models is used to calculate the efficiency scores for some of the DNSPs compared to other DNSPs, the comparison of efficiency scores between DNSPs is not performed on a like-with-like basis" (Evoenergy, 2023a: 11). Evoenergy also notes in this context that the efficiency scores can differ quite considerably across models. The fact that the extent of monotonicity violations, and the estimated efficiency scores, vary between models, Evoenergy argues, calls into question the reliability of the models. This issue is discussed in section 3.1.

(5) Evoenergy appears to argue that a preferred functional specification should be chosen between the Translog (TLG) and Cobb-Douglas (CD) models, rather than averaging the results of these two models. Two criteria relevant to this choice are statistical criteria and economic principles. If a statistical test is relied on to choose between the TLG and CD models, then Evoenergy suggests that the test previously carried out by Quantonomics can be used (Quantonomics, 2022a: 141,146). Economic criteria will require the use of the CD specification if the TLG model gives rise to excessive monotonicity violations, although Evoenergy argues that biased estimates can occur

in the cases where the CD model has not met statistical criteria (Evoenergy, 2023a: 12). The issue of functional specification is discussed in section 3.2.

(6) It is observed by Evoenergy that the estimated output weights derived from the opex cost function model are sensitive to which econometric specification and which sample period is used in estimation. Thus, "the output weights vary significantly between the models" (Evoenergy, 2023a: 14). Since the output weights are used in both the benchmarking roll-forward analysis and the base-step-trend forecasting procedure, the inclusion or exclusion of some of these models "can have a substantial bearing on the outcome of the outcome of these calculations" (Evoenergy, 2023a: 14). Evoenergy is concerned that in the Translog models, the estimated output weights, when averaged by country, differ significantly between Australia, New Zealand and Ontario DNSPs and are strongly influenced by the latter. Hence, the sample average output weights "do not necessarily reflect well the output weights for Australian DNSPs" (Evoenergy, 2023a: 16). A third concern about output weight estimates raised by Evoenergy is that a comparison of estimates obtained using the data sample 2006 to 2021 with those using the 2012 to 2021 sample suggests that "the output weights are not stable over time. Hence, the output weights calculated from the historical data may not be representative of the output weights that apply over an upcoming regulatory period" (Evoenergy, 2023a: 16). The issues relating to output weights are discussed in section 3.3.

(7) Evoenergy is concerned that Australian DNSPs account for only 19 per cent of the sample used in the econometric analysis, with most of the data sample being overseas DNSPs, which operate in different circumstances to Australian DNSPs. Evoenergy claims that the 'country dummy variables' do not adequately control for country-specific differences, citing the authority of the Australian Competition Tribunal (2016). It is claimed that the relationship between non-capital inputs and explanatory variables differs for overseas DNSPs and this influences the ability of the models to provide reliable efficiency comparisons between Australian DNSPs (Evoenergy, 2023a: 36). This issue is discussed in section 3.4.

(8) A further criticism that Evoenergy makes of the econometric opex benchmarking analysis is that opex is analysed separately from capex. Consequently, it does not account for efficient opex-capex substitution choices (as distinct from differences in capitalisation rates which are to be addressed following the AER's recent consultation on capitalisation practices) (Evoenergy, 2023a: 35–36). This issue is discussed in section 3.5.

1.2 Key conclusions

The main conclusions of this memo on the foregoing issues are:

- The number of OEFs taken into account by the AER is not unusually small, although there may be other OEFs yet to be considered, which could be tested as part of the benchmarking development program.
- The AER has said that its forward work program for developing the benchmarking framework includes "updating the quantification of material OEFs … and considering whether GSL payments should be included in opex for benchmarking" (AER 2022, 6). This includes improving the vegetation management OEF (AER 2022, 59), noting the AER has developed and implemented a vegetation management OEF in its recent applications of benchmarking to assess the efficiency of base opex (AER 2021, 28–30). Another potential material OEF identified by Sapere-Merz related to network topology, which is not explicitly mentioned as being included in the benchmarking development program. In our view, it would be useful to include a consideration of whether such a measure can be quantified adequately when the material OEFs are updated.
- Evoenergy's argument that by not adjusting variables prior to, or incorporating them within, the econometric analysis, a bias is introduced into the efficiency scores ultimately determined, is not adequately established. Without any supporting analysis, there is no basis for the claim that its proposed approach would yield a material difference to the AER's approach of making a post-modelling adjustment to the efficiency score. Furthermore, it is not demonstrated that the proposed approach is feasible. We have observed that:
    - Making adjustment to variables prior to the econometric analysis would require estimates of the actual effects of each OEF for each DNSP *in each year of the sample period*. This would be a challenging task, and would be subject to significant estimation error.
    - At present it is not feasible to include many of the OEFs within the econometric modelling due to lack of comparable data for overseas DNSPs.
- We have investigated Evoenergy's claim that monotonicity violations are primarily caused by outliers (see Appendix A). Although other analyses may be possible, the analysis presented in Appendix A suggests there is little correlation between outliers and monotonicity violations, which does not support Evoenergy's contention. Economic Insights has previously suggested that monotonicity violations in the nonlinear models are likely driven by observations with high leverage in the extreme regions of the sample where the data is thinner. Lowry and Getachew (2009) also suggest that multicollinearity among the outputs may also be an important factor. These considerations suggest that the causes of monotonicity violations are likely to be more complex than suggested by Evoenergy.

- We consider it appropriate to exclude the results of models that do not adequately satisfy the requirements from economic theory, as is the case when there are monotonicity violations. In addition, we have previously raised the possibility that, when a TLG model is excluded, in some circumstances it may be desirable to replace the excluded model with a 'hybrid' specification—ie, a constrained TLG model—which does not have excessive monotonicity violations.

- With regard to the general criticism that the degree of variation of efficiency scores across models suggests weaknesses in the models, we show that the variation of efficiency scores between models is actually relatively small. In our view, differences in efficiency score estimates between models merely indicates that the use of different model specifications and sample periods will produce slightly different results and that is precisely why different model specifications and sample periods are used. The AER's practice of averaging output weights and averaging efficiency scores across models should mitigate concerns about the sensitivity of results to model or period chosen.

- The AER's practice of combining the efficiency measures of different models is entirely legitimate if the efficiency measures being combined consistently measure the same inputs and outputs, and the models being averaged are valid and approximately equally performing. This is the case with the AER's econometric opex cost function Translog and Cobb-Douglas models. The averaging of results from these models is consistent with ameliorating concerns about the sensitivity of results to the specific model estimated. Evoenergy's proposal that a single preferred econometric specification be chosen does not appear to be consistent with its concerns about the sensitivity of key parameter estimates to the chosen model specification or sample period.

- We do not accept Evoenergy's apparent argument that the TLG model is to be preferred over the CD model, based on the results of the Wald test of the joint significance of the higher-order terms, because:
    - There are other criteria of model selection to be considered, including goodness-of-fit. Because goodness-of-fit measures penalise loss of degrees of freedom (ie, reward parsimony) the higher order terms can be jointly significant while at the same time, the fit is not improved. This has been shown to be the case in relation to the TLG and CD models.
    - It is difficult to reconcile Evoenergy's apparent argument that the TLG model is to be preferred over the CD model, based on the joint significance of the higher-order terms, with its view that the varying rates of monotonicity violations in the TLG models when applied to different periods casts doubt on the reliability of all of the TLG opex cost function models, not just those with monotonicity violations.

  - In our view, at the present time, there is insufficient basis for entirely excluding the CD models or the TLG models, since they each have strengths and weaknesses when more than one criterion is used in evaluating them.

- On the question of whether the output weights used in calculating the trend component of the base-step-trend forecast should be based on the same average output elasticities or on the average values for Australian DNSPs, as Economic Insights has previously said, whilst there is good economic justification for using the full sample mean Translog output weights, we are relatively indifferent to whether these or the average for Australian DNSPs is used. There may also be some advantages to using the Australian sample average weights.

- For the purpose of combining output forecasts into an output index, the AER uses the output weights derived from the longer sample period. This is a reasonable approach because:
  - There is no reason to suppose that the weights derived from the shorter sample period provide a better basis for forecasting than those from the long sample period; and
  - The longer sample output weights are estimated with more precision.

- Evoenergy's claims that the underlying cost-output relationship is different for overseas DNSPs are not convincing due to a lack of supporting evidence. On the contrary, there is evidence that when a larger weight is given to observations for Australian DNSPs in the regression analysis, so that the relative weight attributed to overseas DNSPs is reduced, there is only a small effect on estimated efficiency scores. This supports the conclusion that the importance of New Zealand and Ontarian DNSPs in the sample does not render the efficiency results unreliable for Australian DNSPs.

- We accept the AER's conclusion that opex/capital substitutability is to some extent indirectly taken into account in our econometric opex cost function models, although the extent to which this is the case is unknown (AER 2023b, 5). This is because of the high correlation of the outputs in this modelling and a capital input variable. If the omitted capital input is closely correlated with the outputs, then to some extent it may be accounted for in the measurement of opex efficiency through the econometric opex cost function models. At this stage we have no information to suggest that substitutability between opex and capital inputs is substantial, and therefore no basis for concluding that the omission of long-run opex/capital substitution effects has a material effect on the reliability of opex cost function efficiency scores. However, it would be desirable to explore this question empirically by estimating a long-run opex cost function using a price ratio between capital and non-capital inputs. This is a substantial task that could form part of the opex cost function development program.

- The AER's adjustment to the efficiency scores by dividing by an efficiency target of 0.75 (rather than 1.0), acknowledges that there are limitations in the benchmarking modelling and provides generous margin for data errors and modelling uncertainties to account for the benchmarking limitations. This finding is inconsistent with Evoenergy's claim that the AER does not adequately account for the limitations of the benchmarking analysis.

2 Accounting for the effects of operating environment factors

2.1 Introduction

Utilities tend to operate in discrete geographical areas, and features of the geographical location, including topography, characteristics of the urban areas supplied (e.g., density) and climate in those locations, may all have an important influence on observed productivity. It is common benchmarking practice to control for the most important OEFs if they are beyond management control.[1] The most important OEFs are those which most influence costs and differ substantially between DNSPs.

As set out below, different procedures can be used to control for the influences of OEFs, such as:

(1) Ex-ante (ie, before the econometric analysis) adjustment of data.

(2) In the econometric analysis, either by including OEFs as variables in the model alongside the outputs or by dividing the data sample into subsamples of similar DNSPs.

(3) After the econometric analysis by adjusting estimated efficiency measures for the influence of OEFs. This may involve quantifying the effect of the OEF using information external to the data sample to determine the appropriate adjustment or carrying out a 'second-stage' regression analysis in which the efficiency scores are dependent variables and OEFs are the independent variables.

There are advantages and disadvantages with each of these approaches for controlling OEFs, and some approaches may not be practical for some OEFs. What is generally needed in benchmarking exercises is the most fit-for-purpose approach in the relevant circumstance.

In section 2.2, background is provided in relation to the range of OEFs to be controlled for across benchmarking approaches and relevant considerations for doing so. Section 2.2 also

[1] By 'beyond management control' we mean that the operating environment factors are exogenous for the firm. Management can still make choices in how to deal with operating environment factors (which may be more or less effective), but these responses generally require resources to implement, so that differences in operating environments can affect the observed comparative productivity and cost efficiency of firms even when action is taken to mitigate their effects. The effects of operating environment factors are an empirical question.

provides background in relation to the AER's approach to incorporating OEF adjustments into its benchmarking.

Evoenergy raises two distinct points in relation to controlling for the effects of OEFs in AER's opex benchmarking.

- The first criticism relates to the range of OEFs controlled for in the benchmarking models, which Evoenergy argues is insufficient to account for the large differences in the operating environments of DNSPs. This is discussed in section 2.3.
- A second concern of Evoenergy is with the practice of making ex-post adjustments to efficiency scores derived from econometric modelling to account for the effects of specific OEFs. It is argued that the effects of different OEFs should be removed from cost data before carrying out econometric analysis. Otherwise, the efficiency ranking of DNSPs may be unreliable, and the reference group comprising the five most efficient DNSPs may not actually include the correct DNSPs. This is discussed in section 2.4.

The types of OEFs to control for and the procedures for doing so are not independent since some procedures may be better suited for some OEFs depending on the quality of information available on the OEF and its effect on costs. This topic is discussed in section 2.4.

2.2 Background on the Range of OEFs to be controlled-for

Section 2.1.1 describes some types of OEFs sometimes accounted for in energy network benchmarking studies and some of the issues that can arise when controlling for OEFs. Section 2.1.2 outlines some background to the OEFs currently included in the AER's DNSP opex benchmarking analysis, including how they are incorporated.

2.2.1 Controlling for OEFs in benchmarking analysis

The aim of making like-for-like comparisons in benchmarking studies supports taking OEFs into account. Since they may impose constraints on the ability to achieve cost-efficiency improvements, regulatory targets informed by benchmarking may be inappropriate if they are not taken into consideration. That said, data on OEFs and their effects on costs tends to be more difficult or costly to obtain than data on outputs and inputs, which will often constrain the degree to which they can be included in a benchmarking study. In a regulatory context, benchmarking needs to rely on objective and auditable data, and this can constrain the information that can inform the regulator on the effects of OEFs on opex.

Among the various OEFs that are sometimes included in energy network benchmarking efficiency studies via the different procedures outlined above are:

- *Terrain and land use*, including the extent of urbanisation and rural land use type. These factors can affect the types of infrastructure and its maintenance requirements. For example, there are specific factors affecting the cost of maintenance in remote areas (such as travel distances) and in heavily built-up areas (relating to congestion). Vegetation management activities, which represent a substantial component of opex, are affected by terrain and land use (eg, in forested areas).

- *Climate conditions* such as storms, high winds and extreme heat or extreme cold can have a material impact on network operations. Some benchmarking studies have found weather to be a decisive factor in explaining observed efficiency differences between energy networks (Yu et al., 2009).
- *Network configuration* is influenced by terrain and urbanisation, but may also be influenced by historical development patterns or the concentration or dispersion of demand centres. Various factors will influence the design of networks which in turn may affect their operating and maintenance costs. For example, in a productivity benchmarking study of Columbian electricity transmission networks, Cadena et al (2009) employ a network complexity index as an OEF.
- *Jurisdictional regulations and standards,* such as undergrounding requirements, network reliability standards and environment protection requirements.
- *The average age of assets*, which is within management control in the long-run, but not in the short-run. For long-lived assets, their average age may depend on the historical growth rates and other patterns of urban development in the supply area. Maintenance requirements can increase for older assets (Diewert, 2009: 12).
- *Market characteristics,* such as the rate of market growth or the customer mix have been included in some studies as OEFs. Since utilities usually have certain obligations to supply the demands within their area, these characteristics may to a substantial degree be beyond management control.

The viability of including specific OEFs in opex benchmarking analysis depends on matters such as:

- whether they can be adequately measured without undue cost. Quantifying the effects of some OEFs on opex may depend on compiling detailed information about the regions supplied by DNSPs and the locations of their existing assets, and may require supplementary analysis of this data, all of which can be costly;
- whether they overlap with other measures already included in the benchmarking analysis via one of the procedures outlined above. As discussed in section 2.4, care is needed to deal with the potential for overlap when OEF adjustments are made before or after regression analysis. This problem does not arise if OEFs are included in regression analysis, although degrees of freedom and other statistical imitations may constrain the extent to which OEFs can be practically incorporated in the opex regression model;

- whether they have a material differential impact on the opex of DNSPs included in the data sample. The influence of factors that do not individually have a substantial differential impact on DNSPs' opex may be regarded as essentially random for the purposes of econometric analysis;
- whether the estimated effects of OEFs on costs are consistent with efficiency. Although OEFs are beyond management control, the way that the business responds to these effects may be efficient or inefficient. Only the effect on efficient costs is to be included. This issue is important for some methods of adjusting for OEFs as discussed in section 2.4.

Because of the granularity of information often needed to measure and assess the effects of OEFs, a regulator may need to rely on regulated businesses for some of the information. Here, adverse incentives can arise due to information asymmetry on the effect of OEFs on opex. For example, if OEF adjustments are to be proposed by DNSPs, and evaluated by the regulator, then as emphasised by Hawdon, there will be regulatory judgements to make about which types of factors to allow for.

> 'The environment in which the gas industry functions varies considerably from country to country, in terms of the terrain over which gas is transported, the geographic density of customers, and their economic characteristics. While this is easily recognized, treatment of such individual circumstances can be affected by strategic considerations. Producers have an interest in stressing the uniqueness of the conditions of supply since regulatory concessions often flow from such recognition. Any such concessions may however be welfare reducing as they remove pressure on producers to improve efficiency in the absence of properly functioning competitive markets. This creates a presumption against including measures of uniqueness where it is desired to assess relative performance unless a priori considerations are overwhelming.' (Hawdon, 2003)

Given the foregoing practical constraints, most benchmarking studies that control for OEFs concentrate on a relatively small number of effects that are considered to have the most significant effect on cost, and which vary the most across DNSPs.

2.2.2 Background on OEF adjustment in the AER's DNSP benchmarking reviews

Of the three methods of controlling for OEFs discussed in the introduction to section 2, the AER predominantly uses methods (2) and (3).[2]

In the early development of the benchmarking framework, Economic Insights (2013a, 2013b) commented on the OEFs that could be included. It was noted that customer density is an

[2] For example, the AER controls for differences in the classification of standard control services (SCS) between jurisdictions by using the narrower concept of 'network services' for opex benchmarking. If differences in the coverage of SCS between DNSPs are due to differences in jurisdictional regulations, this can be considered as an ex-ante OEF adjustment; ie, method (1).

important OEF because a DNSP with lower customer density will generally require more poles and wires to reach its customers than a DNSP with higher customer density. However, when both customer numbers and line length are included as separate outputs, the need to include a customer density OEF is greatly reduced. Similar reasoning applies to energy density when both maximum demand and customer numbers are included as outputs. In addition to these effects, the proportion of circuit length that is underground has also been included in the econometric modelling as an OEF, since underground cables generally have higher capital costs and lower maintenance costs than overhead lines. These are the OEFs that are currently controlled for within the econometric analysis; ie, method (2).

With regard to other OEFs, Economic Insights noted that terrain can significantly affect a DNSP's costs and recommended that a terrain OEF be included in the econometric analysis, but there was at that time "a dearth of terrain summary indicators" (Economic Insights, 2013a: 36). After consultation, it was proposed that ex-post OEFs for bushfire risk and vegetation encroachment be developed. Economic Insights (2013b: 20) also suggested that since reliability is included as an output it is important to control for severe weather-related impacts. These are among the OEFs which the AER controls for using method (3).

The AER formulates DNSP-specific OEF adjustments that are made after the econometric analysis in the context of regulatory reviews, rather than at the time of conducting its annual benchmarking reviews. This has influenced its approach to controlling for OEFs. Prior to the Sapere-Merz (2018) review of OEFs, the AER made post-modelling adjustment for the following OEFs:

- up to 9 material OEFs (depending on the jurisdiction);[3] and
- the combined effect of approximately 20 other OEFs that are individually immaterial but are material in aggregate.[4]

In 2018, the AER decided to engage Sapere-Merz to advise on the selection and quantification of material OEFs and to move away from including immaterial OEFs. The inclusion of immaterial OEFs in its 2015 decisions for NSW/ACT distribution businesses reflected a decision to adopt a cautious approach in the context of the initial application of benchmarking based on more limited data. After the review conducted in 2018, it decided that the immaterial

[3] Subtransmission intensiveness, division of responsibility for vegetation management, extreme weather events, cyclones, bushfire risk and regulations, network accessibility, taxes and levies, OH&S regulations, termite exposure and license conditions.

[4] Including asset age, building regulations, mining boom cost imposts, corrosive environments, cultural heritage, environmental regulations, environmental variability, grounding conditions, proportion of 11Kv and 22Kv lines, rainfall and humidity, skills required by different DNSPs, solar uptake, topography, traffic management, bushfires, capitalisation practices, private power poles, and transformer capacity owned by customers (Frontier Economics, 2019). See: the 2015 Ergon decision (AER 2015, 165–71).

OEFs now represented an overly conservative estimate of the impact of OEFs and their use likely overestimated the magnitude of the operating environment differences between DNSPs.

In its 2018 review of OEFs for the AER, Sapere-Merz was retained to present advice about material differences in OEFs, to identify the most material operating environment factors driving apparent differences in estimated productivity between DNSPs and to quantify the likely effect of each factor on operating costs in the prevailing conditions (Sapere-Merz 2018, 1). Sapere-Merz examined a range of potentially material OEFs, including those suggested by the businesses as likely to be the most material, reaching conclusions on their measurement and materiality. It found that some previously quantified OEF candidates did not meet the OEF criteria when considered across the full set of DNSPs, and its quantification of some OEFs differed from the AER's previous calculations. The material OEFs identified and quantified by Sapere-Merz were:

- Sub-transmission and licence conditions
- Taxes and levies
- Termites
- Extreme weather - cyclones
- Backyard reticulation.

In addition, vegetation management was also identified as a material OEF, but Sapere-Merz was unable to quantify a vegetation OEF at that time. Sapere-Merz also noted that two OEF candidates which were not assessed could be material:

- network topology and
- Guaranteed Service Level payments.

These have not been included to date. The AER has included an additional OEF for differences in occupational health and safety regulations and costs between jurisdictions. Furthermore, in its subsequent decisions, the AER has retained two specific vegetation management OEFs:

- Bushfire risk — the effects on opex of variations in mandated standards of bushfire mitigation activities, specifically the bushfire regulations in Victoria;
- Division of responsibility — the differences in opex between DNSPs due to differences in the division of responsibility for vegetation clearance between the networks, local councils, road authorities and landowners.

The AER has said that its benchmarking development program includes "updating the quantification of material OEFs … and considering whether GSL payments should be included in opex for benchmarking" (AER 2022, 6). This includes improving the vegetation management OEF (AER 2022, 59), noting as above the AER has developed and implemented a vegetation management OEF in its recent applications of benchmarking to assess the efficiency of base opex (AER 2021, 28–30). Another potential material OEF identified by Sapere-Merz related to network topology, which is not explicitly mentioned as being included in the benchmarking development program. In our view, it would be useful to include a consideration of whether such measures can be quantified adequately when the material OEFs are updated.

2.3 Evoenergy's Criticisms – Range of OEFs Controlled For

Evoenergy criticised the range of OEFs that are controlled for in the benchmarking models as insufficient.

> 'The limited number of OEFs recognised by the AER and incorporated into the benchmarking analysis is inadequate to allow proper, like-with-like comparisons between DNSPs that would produce reliable estimates of efficient opex' (Evoenergy, 2023a: 6).

> 'In Evoenergy's view, the small number of opex drivers included within the AER's benchmarking models, and the very limited number of OEFs the AER takes into account in its benchmarking analysis, are inadequate to account properly for the vast differences between DNSPs. This makes the efficiency estimates produced by the AER's models unreliable. …' (Evoenergy 2023b, 35).

As discussed in the previous section, the AER's benchmarking takes account of several OEFs within the econometric analysis, including customer density and the extent of undergrounding of lines. In addition, jurisdictional dummy variables account for all static OEF differences between overseas and Australian DNSPs included in the sample. The AER also makes ex-post adjustments for eight other OEFs as outlined in section 2.2.2.

By way of comparison, in Ofgem's electricity distribution benchmarking in both RIIO-1 and RIIO-2 determinations, it makes ex-ante data adjustments for regional labour cost differences between three regions (London, the South-East, and elsewhere). This is the only OEF adjustment made.[5] In addition, some company-specific adjustments are made in limited circumstances relating to unique characteristics or circumstances of some DNSPs (Ofgem, 2022).

In a 2017 report on benchmarking methods for the Netherlands Authority for Consumers and Markets, Economic Insights surveyed 13 benchmarking studies of the electricity transmission sector. In total these studies used 42 output variables (3.2 per study on average), 27 input

[5] Its three totex models each include a Composite Scale Variable (two different definitions are used) and time trend variables, and one of the three totex models includes demand drivers via a Composite Low-Carbon Technology (LCT) uptake variable based on an equal weighting of the cumulative number of heat-pump (HP) installations and the cumulative size of electric vehicle (EV) chargers. No OEFs are used in these totex regression models.

variables (2.1 per study), and 19 OEFs (1.5 per study on average) (Economic Insights, 2017: 33–37).

These observations suggest that the number of OEFs taken into account by the AER in its benchmarking analysis (10 or more using different procedures outlined above) is not unusually small. Nevertheless, there may be other OEFs yet to be considered, which could be tested as part of the benchmarking development program. Section 2.2.2 has a brief outline of the AER's statements relating to OEFs in the benchmarking development program.

2.4 Evoenergy's Criticisms – Method of Controlling for the effects of OEFs

Evoenergy also criticises the AER's method of controlling for most OEFs by making ex-post adjustments to efficiency scores derived from econometric modelling. Evoenergy argues this is likely to produce unreliable estimates of efficiency for individual DNSPs and potentially misidentify reference DNSPs. In particular, it argued the reference group of DNSPs obtained from the econometric modelling may be incorrect.

> 'The Australia Competition Tribunal noted in 2016 that it would be preferable for the data used in the benchmarking analysis to be adjusted and normalised to improve comparability before applying it to a benchmarking model to determine the efficiency scores (i.e., an *ex-ante* approach to adjusting for OEFs)' (Evoenergy, 2022: 3).

In the context of considerations around how to address capitalisation differences between DNSPs, Evoenergy recommended that the AER's benchmarking development program should examine whether OEF adjustments can be implemented using an ex-ante method. In its submissions to the AER's consultation on its Capitalisation Guidance Note, Evoenergy noted that "making ex-post OEF adjustments to DNSPs' efficiency scores after modelling (which uses unadjusted, non-comparable data) has been undertaken to determine these scores is problematic" because the DNSPs identified as 'reference' DNSPs is based on benchmarking models that have not properly accounted for all of the inherent differences between DNSPs. Evoenergy urged the AER to "undertake further work in the future to determine whether all OEF adjustments could be implemented in an ex-ante fashion" (Evoenergy, 2022, 2023c). In its regulatory proposal, Evoenergy states that "The purpose of OEF adjustments is to make DNSPs more comparable to one another, *before* assessing their efficiency, by accounting for differences in factors that drive opex but are unrelated to efficiency" (Evoenergy, 2023a: 23 italics added).

As outlined at the start of section 2, different procedures can be used to control for the influences of OEFs. As discussed in section 2.1.2, the AER predominantly uses the second and third of the following three alternative methods:

(1) Ex-ante (ie, before the econometric analysis) adjustment of data: This involves adjusting the opex or output measures for the effects of the OEFs, thereby making the resulting measures more closely comparable. This approach typically relies on

information external to the data sample, such as engineering knowledge, to determine the appropriate adjustment.

(2) In the econometric analysis, particularly by including OEFs as variables in the model alongside the outputs. In the AER's benchmarking framework, this is currently done with: (a) the inclusion of a variable for the share of circuit length that is underground; (b) the inclusion of country dummy variables which account for various static differences between the jurisdictions, including differences in opex coverage and OEFs; and (c) differences in customer density are also accounted for by the output specification.

(3) After the econometric analysis by adjusting estimated efficiency measures for the influence of OEFs. This may involve quantifying the effect of the OEF using information external to the data sample to determine the appropriate adjustment (which is the AER's approach) or carrying out a 'second-stage' regression analysis in which the efficiency scores are dependent variables and OEFs are the independent variables.

The different methods have strengths and weaknesses for different types of measurement approaches, and therefore different methods may be most suitable for controlling for different OEFs. These methods are discussed in turn below, including a response to Evoenergy's arguments.

2.4.1 Ex-ante adjustment for OEFs

Ex ante adjustments are commonly made to ensure that data is measured on a like-for-like basis. Ex-ante adjustments for OEFs require adjustments to the data to remove the differential effect of an OEF, before other analysis is undertaken. One example of ex-ante adjustment of data for OEFs is the process undertaken by Agrell and Bogetoft (2009), in which electricity transmission businesses were given the opportunity to submit adjustments they believe should be made to the data to put them on a like-for-like basis. These proposals were assessed on a case-by-case basis. This process clearly has problematic incentives in a regulatory context for reasons given by Hawdon as quoted above.

In recent European energy transmission benchmarking, output variables were adjusted ex-ante for routing complexity due to land use and terrain, and for different ground and subsurface conditions (Sumicsid and CEER 2019). This can be a highly data-intensive exercise, particularly if such adjustments are made at the individual asset level, requiring granular information on assets and on the operating and maintenance costs of those assets. It also relies heavily on engineering knowledge to quantify the effects of different conditions on costs. The effect of factors such as terrain or climate on network design and maintenance cost is a complex question. The extent of the required engineering analysis and data management means that such adjustments can be costly to carry out. Moreover, the Sumicsid study was subject to criticism in relation to its ex-ante OEF adjustments (Oxera, 2020). Because it relied on granular data for the benchmarked businesses, there were confidentiality restrictions on disclosure of the detailed ex-ante data adjustments made. This and other disclosure issues relating to assumptions made in quantifying the effects of the OEFs on cost items, it was claimed that the OEF adjustments could not be adequately examined or validated.

Another difficulty that can arise with ex-ante adjustment in circumstances where an OEF has broad and complex effects that are not simple to quantify outside the regression model is that it can be difficult to determine whether the adjustments made accurately reflect the effect of the OEF on costs.

These observations suggest that ex-ante adjustment of data is not always desirable. However, it is a practical approach when the cost variable can be adjusted for factors that can be objectively quantified.[6]

2.4.2 Adjustments within the econometric analysis

The main approach to controlling for OEFs within a regression model is by including them as additional variables in the econometric analysis. Alternatively, if a relevant set of OEFs can be used to separate DNSPs into a small number of sub-groups representing those with similar network configurations, then data analysis could be carried out separately for each sub-group. This approach depends on having a large data sample, and with the data sample used by the AER, this approach may not be feasible. Usually, a similar effect can be achieved by using dummy variables for the same categories within a full sample analysis. Many econometric benchmarking studies include OEFs in the regression equation, although they usually include only a few such variables. For example, Orea and Jamasb (2017) estimate a total cost function for Norway electricity distribution networks,[7] and include as OEFs two weather variables (average wind speed and average wind exposure) and a geographic variable (average distance of main supply areas to coast). These OEFs mainly capture the effects of wind and salinity on accelerating the deterioration of network assets.

As stated above, the AER's econometric opex cost analysis takes account of some key OEFs, including customer density, undergrounding and static jurisdictional differences.

The approach of including OEFs in the regression model can have advantages over some of the alternative methods because:

[6] The AER controls for differences in the definition of standard control services (SCS) between jurisdictions by using the narrower concept of 'network services'. To the extent that differences in the coverage of SCS between DNSPs are due to differences in jurisdictional regulations, this might be viewed as an ex-ante OEF adjustment.

[7] Total cost here includes the utility total cost of supply plus the consumer cost of energy not supplied. The data sample had 957 observations on an unbalanced panel for the years 2004 to 2011.

- The effect of an OEF on opex can be determined statistically, rather than quantified based on non-sample information. This provides greater confidence that the claimed effect of the OEF exists and an objective basis for its quantification.
- The statistical significance of the effect can be tested. In principle, this can assist to ensure that only the most important OEFs are controlled for.[8]
- The potential for overlap and hence double-counting of the effects of an OEF is reduced because regression coefficients represent the partial effects of each explanatory variable, given the other variables in the model.

However, there are some limitations to including OEFs within regression models, including the loss of degrees of freedom associated with expanding the number of explanatory variables; and the potential for multicollinearity between OEFs. These issues might arise if there are many OEFs to include in a model. Both of these issues can be mitigated by using Principal Components Analysis to reduce the set of OEFs into a smaller number of uncorrelated variables which contain most of the variation in the original variables (Economic Insights, 2017). Another limitation, which arises in the AER's DNSP benchmarking exercise, is where information on the OEF is only available for some of the DNSPs in the sample, namely the Australian DNSPs. It may then be infeasible to use this approach. Economic Insights states that lack of consistent data availability "across the entire three country sample precluded the explicit inclusion of additional operating environment factors" in the opex cost function model (Economic Insights, 2015: 66).

This method would generally be preferred when adjusting for broader OEFs with effects on opex that would otherwise be difficult to quantify, and when there is a sufficiently large sample and consistent data on the OEF for all DNSPs in the sample. It may not be a preferred method of controlling for those OEFs whose effects can be more accurately estimated using non-sample data, and where potential issues of overlap can be adequately avoided.

2.4.3 Ex-post OEF adjustments

Ex-post adjustment of efficiency scores is quite common in efficiency analysis. Often it takes the form of a second-stage econometric analysis,[9] but in a regulatory context, it may involve

[8] If there is a large number of relevant OEFs or they are highly correlated with each other, then Principal Components Analysis can be used to reduce the number of variables to be included in the econometric model. Alternatively, non-sample information might be used to devise weights for constructing an index of a certain group of OEFs.

[9] Second-stage econometric analysis of the relationship between efficiency scores and OEFs is often used after data envelopment analysis (DEA) because of limitations to including OEFs within DEA. This method is not useful in the econometric analysis of efficiency because OEF variables could be included at the first stage. If the OEF variables are only available for Australian DNSPs, a second-stage regression analysis of their efficiency scores against the OEFs would only have 13 efficiency scores as observations on the dependent variable, which would be too few to obtain reliable estimates of the OEFs.

making adjustments to efficiency scores to allow for certain DNSP-specific effects, as the AER does. The AER approach to adjusting for material OEFs not included in the econometric models involves:

- Quantifying the effect of the OEF on each DNSP's *efficient* cost;
- Expressing the effect of the OEF on opex in percentage terms relative to its effect on the opex of the reference DNSP (ie, the average of the group of efficient DNSPs);
- Aggregating these percentage effects over all OEFs;
- Adjusting the efficiency target of each inefficient DNSP for the aggregate percentage effect of the OEFs.

This approach is taken primarily because information on the quantity of the OEF and its effect on the opex of each DNSP is only available for Australian DNSPs. It would be impractical to obtain comparable information for overseas DNSPs. This procedure only requires establishing the efficient effect of the OEF on opex at a point in time relevant to the base year used in a particular opex decision.

Evoenergy's proposal that the OEF adjustments should be made ex-ante would require that the actual effects of these OEFs on opex would need to be estimated for each DNSP in *each year of the sample period*. This is because opex for each Australian DNSP in each year of the sample would need to be adjusted to exclude the effect of the OEFs. This would be a much more challenging task than that undertaken by Sapere-Merz, which was estimate to the efficient effect of the OEF on opex at a point in time, and for the most part, the same percentage effect has then been carried forward. Attempting to adjust opex for OEFs in each year would likely be subject to considerable estimation error. This may in turn compromise the reliability of the econometric analysis. This is an important practical problem not given enough recognition by Evoenergy in its submission.

The OEFs for which the AER makes ex-post adjustments are such that their effects that can be adequately quantified separately for DNSPs in a reasonably objective way. This approach is only used for those OEFs where data is not universally and consistently available across countries, but is available for Australian DNSPs. Hence, these OEFs cannot be included in the econometric analysis but can be done as an adjustment when comparing Australian DNSP performance to the most efficient Australian firms. Where consistent data on an OEF is available for DNSPs in New Zealand, Ontario and Australia, the AER will include it in the econometric analysis where appropriate. Hence, the AER adopts a combination of approaches.

Evoenergy's argument that the reference group of DNSPs may be incorrectly selected when OEF adjustments are undertaken ex-post seems to relate, not so much to whether the OEF adjustment is made ex-ante or ex-post, but to whether the reference group is selected before or after the OEF adjustments are considered. The AER normalises the efficiency score (ES) of a DNSP by dividing it by the adjusted comparison point (ACP). $ACP = 0.75/(1 + OEF\%)$, where $OEF\%$ is the aggregated percentage effect of the OEFs relative to the reference group average. Mathematically, this is equivalent to calculating an adjusted efficiency score using, $ES^A = ES \times (1 + OEF\%)/0.75$. The ranking could then be redone based on the adjusted efficiency score.[10] However, while noting that this appears to be feasible, it needs to be emphasised that if there is no change in the group of reference firms (which will usually be the case because of the small size of OEF adjustments relative to differences in inefficiency between DNSPs) an inefficient firm's adjusted efficiency score will be unaffected by this additional step. If there *is* a change in the group of reference firms, the effect on the adjusted efficiency score of the inefficient firm is likely to be very small.

Evoenergy's argument that not adjusting variables prior to the econometric analysis means unreliable estimates of efficiency scores for individual DNSPs are obtained, is not adequately established. Without any supporting analysis, there is no basis for the claim that its proposed approach would yield a material difference to the AER's approach of making a post-modelling adjustment to the efficiency score. Furthermore, it is not demonstrated that the proposed approach of making ex-ante adjustments to opex is feasible. We have observed that:

- Making an OEF adjustment to variables prior to the econometric analysis would require estimates of the actual effects of each OEF for each DNSP in each historical year of the sample period. This would be a challenging task, and would be likely subject to significant estimation error.
- At present it is not feasible to include within the econometric modelling many of the OEFs currently accounted for ex-post due to lack of comparable data for overseas DNSPs.

3 Limitations of the econometric opex cost function

This section discusses specific modelling issues relating to the econometric opex cost function used to derive estimates of opex efficiency.

- The implications of monotonicity violations are discussed in section 3.1.
- Selecting a preferred model versus model averaging is discussed in section 3.2.
- Issues relating to differences in output weights between models and between sub-samples are discussed in section 3.3.

[10] If the ranking has altered then a second iteration is needed to restate the OEF% amounts by subtracting by a constant; namely, the average OEF% for the new set of five reference firms (call this constant A). The adjusted efficiency score (ES^A) can then be recalculated. No further iteration is needed because all the adjusted efficiency scores are rescaled by approximately the same proportion at this second step; ie, by (1 – A).

- The use of overseas comparators in the data sample is discussed in section 3.4.
- Accounting for efficient opex-capex substitution choices is discussed in section 3.5.

3.1 Monotonicity violations

Evoenergy has noted that the estimates of average efficiency are affected by the removal of those models for which there are monotonicity violations. While the monotonicity violations are limited in the long-period model, there are a considerable number of monotonicity violations in the short-period models. Evoenergy claims that:

(a) the approaches presented in Quantonomics (2022b) have not resolved the problem.

(b) the monotonicity issues are driven by outliers in the data and cannot be addressed. The "the only real solution to the problem of monotonicity violations is to view the estimates from those models with caution and scepticism" (Evoenergy, 2023a: 36);

(c) because "the monotonicity violations occur over some benchmarking periods but not others", this calls "into question the reliability of the models even in those periods where no monotonicity violations are apparent" (Evoenergy, 2023a: 36). Evoenergy is likewise concerned that the efficiency scores can differ quite considerably across models (Evoenergy 2023b, 11).

Issue (b) is addressed in section 3.1.1. Issue (a) relating to the approaches proposed by Quantonomics (2022b) and further approaches are discussed in section 3.1.2. Issue (c) is discussed in section 3.1.3.

3.1.1 Monotonicity violations and outliers

Evoenergy states that "the underlying cause of the monotonicity violations are the outliers in the benchmarking dataset, and the extent to which the translog models must be 'flexed' in order to fit these outlying data well" (Evoenergy, 2023a: 36).

Influential Observations

Before assessing this claim, it is important to clarify certain econometric concepts. An observation is referred to as 'influential' if removing the observation substantially changes the estimates of coefficients. There are two kinds of observations that have the highest influence on regression model estimates: outliers and observations with high leverage (Chatterjee and Hadi, 1986). An outlier is an observation with a large residual—implying that the dependent-variable value is unusual given the values of the predictor variables. An observation with high leverage is one with a large deviation from the mean of that variable—implying that a predictor variable takes an extreme value. High leverage observations are generally not outliers (Greene, 2012: 140). They are the extreme values or atypical combinations of the explanatory variables, such as very large or very small DNSPs, or those with an atypical mix of outputs relative to their overall scale. Hence, highly influential observations are on thinly populated extremes of the observations around the regression hyperplane.

In the 2021 benchmarking report, Economic Insights (2021b: 7) offered the following thoughts about the causes of excessive undue monotonicity violations:

> The monotonicity violations are most likely the result of the greater flexibility in the TLG functional form in which the edges of the isoquants can 'bend backwards' in places because the data is thinner in those extreme regions and hence the shape of the production surface could be more influenced by a handful of atypical observations.

That is, if the function being fitted is non-linear, then in the extreme regions, its shape will be strongly influenced by a relatively small number of observations which have high leverage. In the quoted statement, Economic Insights appears to give most emphasis to observations of high leverage in explaining excessive monotonicity violations.

Are Excessive Monotonicity Violations Primarily Due to Outliers?

In Appendix A, we examine the claim that monotonicity issues are primarily caused by outliers in the data. We recognise that there is a range of different examinations that can be carried out using different definitions of outliers or techniques. That said, the correlation statistics presented in Appendix A do not support a conclusion that monotonicity violations are primarily caused by outliers.

Conclusions

It is not uncommon for Translog models, due to their flexibility, to give rise to monotonicity violations. Kumbhakar, Wang and Horncastle (2015: 107) suggest "imposing more structure in the estimation process" to address this issue. That is, impose suitable parameter constraints. For example, in benchmarking Ontario DNSPs, Pacific Economic Group (PEG) uses a Translog model, but excludes output interaction terms because "the inclusion of such interaction terms leads to a violation of output regularity [monotonicity], for nearly two-thirds of the observations, which is needed for positive marginal costs" (Lowry and Getachew, 2009: 336). Incidentally, Lowry and Getachew suggest that the monotonicity violations are likely due to "multicollinearity between the output variables and a sample of inadequate size" (Lowry and Getachew, 2009: 336). Economic Insights also states that one approach to mitigating monotonicity problems is to impose selected constraints on the TLG model (by setting the coefficients on higher-order terms to be zero), thereby making its "curvature more 'well behaved' in the face of atypical or outlying observations" (Economic Insights 2021b, 140). This is a general statement encompassing both high leverage and outlier observations without addressing the specific causes of monotonicity violations in the opex benchmarking models. The significance of multicollinearity, noted by PEG, is that in some parts of the sample, a high correlation between outputs causes the partial effect of one output to be close to zero. These observations suggest that the causes of monotonicity are complex, and Evoenergy's suggestion that they are caused by outliers is too simplistic.

3.1.2 The effect of excluding models on model-average efficiency scores

Evoenergy observes that "the efficiency scores can differ quite considerably across models", and consequently, when models with excess monotonicity violations are excluded to calculate the average efficiency score of a business, this can unduly affect the measured efficiency and is inconsistent with like-with-like comparisons (Evoenergy, 2023a: 11).

The current and previous practice of the AER and Economic Insights (and more recently Quantonomics) is to remove a model from the model-average efficiency score calculated for a specific DNSP when there are monotonicity violations for a majority of observations for that DNSP, and to remove a model from the average efficiency scores calculated for all DNSPs when the majority of Australian DNSPs have excessive monotonicity violations. This is based on the belief that the estimates produced by models with excessive monotonicity violations are unreliable, in that they do not meet the economic principle behind monotonicity. Hence, the average efficiency scores calculated when the estimates from invalid models are excluded should be a better overall estimate of efficiency than an average with the invalid estimates included.

That said, it remains a question as to whether there may be systematic differences between the results of the four model specifications such that altering the number of models over which the average is calculated may have an effect on the average efficiency score beyond the effect of removing the inaccuracy associated with the invalid model. This is a difficult question to address because to separate the two effects we would need the results of TLG models that are restricted just enough to satisfy the requirement that it does not have excess monotonicity violations.

It may be useful in the first instance to consider the scale of the possible overall effects of changing the number of models over which efficiency estimates are averaged, while noting the limitations of this assessment. Specifically, the major limitations are that it treats estimates produced by models with monotonicity violations as if they were equal to estimates that would be produced by a TLG model which has been restricted just enough to satisfy the requirement that it does not have excess monotonicity violations. Of course, this need not be the case, which means that the assessment will be inaccurate.

Table 3.1 shows the efficiency scores estimated in the 2022 electricity DNSP benchmarking study for the 2006 to 2021 period. It also compares the average efficiency scores for all four models to the averages obtained when one of the TLG models is excluded. The absolute differences between the average efficiency scores calculated over three models and the averages calculated over four models are expressed in percentage terms.

Table 3.1 Estimated efficiency scores, 2006–2021

	Estimated efficiency scores				*Average efficiency scores*			*% Absolute differences from Average of all four models*	
DNSP	*SFACD*	*SFATLG*	*LSECD*	*LSETLG*	*All four models*	*Excl. LSETLG*	*Excl. SFATLG*	*Excl. LSETLG*	*Excl. SFATLG*
EVO	0.477	0.499	0.480	0.422	0.470	0.486	0.460	3.4%	2.1%
AGD	0.471	0.387	0.475	0.461	0.448	0.444	0.469	0.9%	4.5%
CIT	0.952	0.943	0.930	0.846	0.918	0.941	0.909	2.6%	0.9%
END	0.611	0.557	0.605	0.608	0.595	0.591	0.608	0.7%	2.1%
ENX	0.609	0.538	0.616	0.604	0.592	0.588	0.609	0.7%	3.0%
ERG	0.584	0.664	0.567	0.587	0.601	0.605	0.579	0.7%	3.5%
ESS	0.604	0.658	0.647	0.720	0.657	0.636	0.657	3.2%	0.0%
JEN	0.657	0.683	0.674	0.526	0.635	0.671	0.619	5.7%	2.5%
PCR	0.969	0.973	1.000	1.000	0.985	0.981	0.990	0.5%	0.4%
SAP	0.751	0.768	0.789	0.819	0.782	0.769	0.786	1.6%	0.6%
AND	0.670	0.663	0.728	0.675	0.684	0.687	0.691	0.4%	1.0%
TND	0.793	0.821	0.769	0.718	0.775	0.794	0.760	2.5%	2.0%
UED	0.816	0.783	0.811	0.646	0.764	0.803	0.758	5.1%	0.8%
Avg.	0.689	0.688	0.699	0.664				2.2%	1.8%

Table 3.1 shows that typically the differences between the average efficiency scores calculated over three models and averages calculated over four models are reasonably small, representing about 2 per cent on average. However, there are a few instances where the effect is more marked. For example, the averages for JEN and UED are most substantially affected when the LSETLG model is excluded, and AGD when the SFATLG model is excluded.

We consider it appropriate to exclude the results of models that do not adequately satisfy the requirements of economic theory because the estimates derived from such models are likely to be unreliable.

Restricted TLG specifications

In our view, if a TLG model is excluded due to excessive monotonicity, it would be preferable to replace it with an estimate drawn from a TLG model restricted to satisfy the requirement that it does not have excess monotonicity violations.[11] Alternatively, all TLG models can be replaced by a constrained TLG specification satisfying the monotonicity requirement as done by PEG (see section 3.1.1). This may minimise any effects on averages that may arise from excluding a model.

We have previously raised the possibility that, when a TLG model is excluded, in some circumstances it may be desirable to replace the excluded model with a 'hybrid' specification—ie, a constrained TLG model—which does not have excessive monotonicity violations (Quantonomics, 2022b). Our 2022 memo considered several so-called hybrid models, which are Translog models with constraints imposed by omitting selected higher-order terms. As noted in section 3.1.1 above, PEG has adopted a similar approach of restricted the TLG specification in its benchmarking of Ontario DNSPs.

The 2022 memo adopted the following criteria for evaluation of model validity:

- the extent to which the model reduces monotonicity errors for Australian DNSPs
- joint significance tests of groups of related explanatory variables added to a model
- goodness-of-fit, and
- the meaningful economic interpretation of parameters.

The analysis was not exhaustive; there are a great many alternative sets of parameter constraints that can be applied. Therefore, further work is needed to formulate a preferred hybrid model. In the conclusion of the Quantonomics 2022 memo, we stated:

> "In circumstances where the Base TLG models have too many monotonicity violations, one of these hybrid models could be used in its place when calculating the average of four econometric models. When making such a substitution, a model with the corresponding estimation technique should be used. For example, if the Base LSETLG model has too

[11] The terms 'restricted TLG', 'constrained TLG' and 'hybrid' are here used interchangeably.

> many monotonicity violations, the efficiency score obtained from the LSE version of the hybrid model should be used in its place (eg, LSETLG-H3). An efficiency score from an SFA hybrid model should not be substituted for an efficiency score from the LSETLG model, and vice versa." (Quantonomics, 2022b: 10)

We did not express a view as to which of the hybrid models should be chosen for that purpose, although it should clearly be a model that has relatively few monotonicity violations and does not have excessive monotonicity violations for the DNSP for which it is used. Although it is a short-term solution, it may go some way to addressing the issue raised by Evoenergy.

Going forward, further work is needed to formulate a restricted TLG specification which consistently meets the criteria stated above over different sample periods, and outperforms alternatives when additional data is added to the sample. It may be that a hybrid model can be used more generally in place of the TLG form, as PEG has done.

3.1.3 Monotonicity and reliability

Evoenergy also notes that monotonicity violations "seem to be quite sensitive to the data period used to estimate the models" (Evoenergy, 2023a: 11). Further, the problem of monotonicity violations has not been ameliorated as the dataset has expanded over time (Evoenergy, 2023a: 12). Evoenergy asserts that this variation in the rate of monotonicity violations between different data periods calls into question the reliability of the opex cost function models.

Sensitivity of monotonicity violations to the sample period used refers to both: changes in the sample periods from one benchmarking report to the next by the addition of the most recent year of data; and differences between the full sample period and the shorter sample period over which the models are estimated. With regard to changes in the sample period from one year to the next, recent benchmarking reports indicate that:

- For the TLG models estimated over the full sample period (from 2006) there was a small number of DNSPs with excess monotonicity violations in the 2020 and 2021 reports.[12] However, in the 2022 report, which is most relevant to Evoenergy at the present time, neither of the LSETLG and SFATLG models had any monotonicity violations for Australian DNSPs (Quantonomics, 2022a: 33).
- For the TLG models estimated over the shorter sample period (from 2012) there has consistently been a higher frequency of monotonicity violations. In this context, the problem of excess monotonicity violations has resulted in the exclusion of a significant

[12] In the 2020 benchmarking report, the LSETLG model had no monotonicity violations for any Australian DNSPs, whereas the SFATLG model had excess monotonicity violations for two Australian DNSPs (Economic Insights, 2020b: 34). In the 2021 benchmarking report, the SFATLG model had no monotonicity violations for any Australian DNSPs, whereas the LSETLG model had excess monotonicity violations for three Australian DNSPs (Economic Insights, 2021b: 28).

proportion of Australian DNSPs, and in some instances, one of the models has been excluded altogether.

These observations suggest that the problem of excess monotonicity violations seems to be mainly limited to the models estimated using the shorter sample period. For example, the opex cost function results in Quantonomics (2022a) show that there is a much higher frequency of monotonicity violations in estimates using the shorter sample period from 2012 to 2021, than when using the longer period from 2006 to 2021. This suggests that the frequency of monotonicity violations is likely influenced by sample size. The models estimated using a larger sample are generally more reliable than those estimated using a smaller sample. As observed by Lowry and Getachew cited above, the performance of the Translog specification in relation to output regularity is likely related to the adequacy of sample size combined with multicollinearity between variables in the TLG specification. As discussed in section 3.1.2, in circumstances where the TLG models do not perform well, a constrained TFG specification may be more suitable. This may be relevant to the short sample period.

Evoenergy has not explained how the observation that the models estimated using a larger sample are less susceptible to monotonicity violations leads it to a conclusion that all the TLG models—whether estimated over a shorter or longer period—are unreliable. It has not given reasons to conclude that a model estimated using the long sample period that has few monotonicity violations (and in the case of the 2022 analysis, no excess monotonicity violations for Australian DNSPs) is in some way rendered unreliable because a model estimated over a shorter period yields a higher incidence of monotonicity violations. In our view, this argument by Evoenergy cannot be convincing in the absence of clear reasoning.

Furthermore, Evoenergy's argument that all of the TLG models are unreliable cannot flow over to the CD models, which have no monotonicity violations, and hence its argument against the TLG models directly contradicts its argument relating to model selection (and use of the TLG models) discussed in the next section.

With regard to the general criticism that the degree of variation of efficiency scores across models suggests weaknesses in the models, we would suggest that, on the contrary, the variation of efficiency scores between models is actually relatively small. Figure 3.4 of the 2022 benchmarking analysis report (Quantonomics, 2022a: 35) shows that the efficiency scores estimated using the different models are reasonably consistent for most DNSPs. In our view, differences in efficiency score estimates between models merely indicate that the use of different model specifications and sample periods will produce slightly different results and that is precisely why different model specifications and sample periods are used. Efficiency scores are measured with reference to an efficiency frontier which is estimated using the data sample and technique employed. The reasonable degree of consistency of estimates of efficiency scores between the different models is an indication of their reliability, and not that the results are unreliable, as Evoenergy suggests. Furthermore, the AER's practice of averaging output weights and averaging efficiency scores across models should mitigate concerns about the sensitivity of results to model specification or sample period chosen.

3.2 Model selection

Evoenergy appears to argue that a preferred functional specification should be chosen between the Translog (TLG) and Cobb-Douglas (CD) models, rather than averaging the results of these two models. Based on the statistical tests carried out by Quantonomics, Evoenergy suggests that the TLG function is to be preferred over the CD function. This section first considers, in section 3.2.1, the relative merits of averaging the results of several valid models versus choosing a preferred functional form and relying on that. Then section 3.2.2 examines the question of whether the TLG outperforms the CD model, drawing on a wider range of specification tests.

3.2.1 Model averaging versus choosing a preferred model

The AER's benchmarking method uses four different econometric specifications, in terms of two alternative functional form specifications in combination with two alternative estimation methods that specify and estimate opex cost efficiency. The average of the efficiency scores of these four models (subject to addressing the monotonicity issue discussed in section 3.1) is used by the AER as a preferred measure of efficiency. Similarly, the averages of the output weights derived from these models are also used.

As noted by Farsi et al, an important problem faced by regulators that conduct benchmarking "is the choice among several or legitimate benchmarking models that usually produce different results" (2007: 1). It is usually desirable to use more than one benchmarking technique for the purpose of methodological cross-checking and to perform diversified analysis. Beyond this, there is the question about whether only one 'preferred' model should be relied on or whether the results of several well-performing models should be combined somehow.

Some practitioners argue that a 'preferred model' should be chosen from among them. For example, Haney and Pollitt:

> … there is a question about whether efficiency scores produced by different methods should be combined. Clearly, simply averaging a set of efficiency scores for the same firm (produced for example by DEA, COLS and SFA or different specifications of the same measurement technique) produces a score which itself does not correspond to the result of any one method. It makes more sense to pick the result of one set of estimates, on the basis of the argument that this was the most appropriate method of measuring the efficiency of the sample of firms in question, and consistently use that. (Haney and Pollitt, 2012: 29–30)

A contrasting view is expressed by Agrell and Bogetoft:

> As long as benchmarking scholars cannot clearly rank one method as being superior to another we see no reason the regulator should make that call. It is also not just an 'easy way out' of methodological discussion to apply multiple methods. In fact one can argue that … the simultaneous application of multiple methods puts additional discipline on the model development approach. (2016: 15)

Greene observes that "if we have doubts about which of two model is appropriate, then we might well be convinced to concede the possibility that neither one is really 'the truth'" (Greene, 2012: 181). Some form of model averaging or averaging of model predictions can be a suitable strategy for minimizing the uncertainty regarding the 'true' model specification.

Consistent with these latter views, some regulators formally combine more than one approach. For example, the German regulator has used the maximum of the efficiency measures obtained using DEA and SFA methods (subject to a minimum value of 0.6) (Kuosmanen, 2012). For electricity distribution cost assessment, Ofgem uses a combination of totex and activity-level cost models. The totex and disaggregated modelling streams are each assigned 50 per cent weight, and its three individual totex models receive an equal share of the 50 per cent weight assigned to totex (Ofgem, 2022: 226). Some researchers have proposed more complex algorithms for combining the results of different methods (Azadeh et al., 2009). When the AER's top-down opex benchmarking methods were challenged in 2016, the Australian Competition Tribunal concluded that the AER should use "a broader range of modelling and benchmarking" (Australian Competition Tribunal 2016, at [1227]).

In practice, one model may not entirely dominate the other models. In a regulatory context it is desirable to have some stability of method from study to study. While one model may have marginal claims to superiority in one sample, when the sample is updated, another model may have marginal claims to superiority. Switching from one to another at each update may give rise to an undesirable degree of instability in efficiency assessments.

Having regard to the evidentiary context of regulatory decisions for which benchmarking is being used, it is not clear that picking one model is to be preferred, especially if its claims to superiority are only slightly better than some other alternatives. In these circumstances, a 'consensus' approach drawing on results from several plausible models has merit. As Farsi et al observed, if there is significant uncertainty in inefficiency estimates, e.g. because there is more than one convincing model that yield different results, this "could have important undesired consequences especially because in many cases the efficiency scores are directly used to reward/punish companies through regulation scheme such as price formulas" (Farsi et al., 2007: 13).

The use by regulators of average efficiency scores of more than one benchmarking model can help to ensure that efficiency assessment is not too dependent on one model specification if they have different strengths and weaknesses and potential specification errors. The average of results can reduce the potential impact of specification error. Furthermore, depending on the selection criteria, the selection of a single preferred model could lead to volatility from year to year if the chosen model specification changes from year to year as the data sample is expanded. Averaging the results of more than one plausible model can promote the stability of the estimates.

In summary, the AER's practice of combining the efficiency measures of different models is entirely legitimate if the efficiency measures being combined consistently measure the same inputs and outputs, and the models being averaged are valid and approximately equally performing. The discussion in section 3.2.2 below supports a conclusion that this is the case in relation to the AER's valid econometric opex cost function Translog and Cobb-Douglas models. The averaging of results from these models is consistent with ameliorating concerns about the sensitivity of results to the specific model estimated, which are precisely the concerns Evoenergy raises in relation to output weights, as discussed below. Evoenergy's proposal that a single preferred econometric specification be chosen does not appear to be consistent with its concerns about the sensitivity of key parameter estimates to the chosen model specification or sample period.

3.2.2 Model selection

As discussed, Evoenergy suggests that a single preferred functional form should be chosen based on statistical and economic tests, such as those carried out by Quantonomics. The specific statistical hypothesis tests referred to are the Wald tests of the joint statistical significance of the higher-order output terms in the Translog (TLG) model. Since the Cobb-Douglas (CD) specification is a restricted form of the TLG model when the coefficients on all of these higher-order output terms are all equal to zero, this serves as a test of the TLG versus CD specifications. When the constraints implied by the CD model are rejected against the TLG model (and if the monotonicity conditions are met) then it could be concluded the TLG model should be used and the CD model rejected. Evoenergy observes that results in Quantonomics (2022b) show that in most cases the statistical hypothesis tests support the TLG models over the CD models (Evoenergy, 2023a: 11). In these cases, and if the TLG has no monotonicity violations for the Australian DNSPs, "it would be difficult to justify the use of the Cobb-Douglas model" (Evoenergy, 2023a: 12). They conclude that in most cases, the TLG model should be preferred over the CD model.

In the memo 'Opex cost function development' (Quantonomics, 2022b), which explored options to deal with the issue of monotonicity violations in the TLG models, several criteria were considered for evaluating alternative models, and one was joint significance tests of groups of related explanatory variables added to a model. Others included goodness-of-fit and the meaningful economic interpretation of parameters. The use of a consistent measure of fit is made more difficult because the LSE model used in opex cost benchmarking is based on Stata's *xtpcse* command for the panel-corrected standard errors model, which is not a maximum likelihood method. Hence, common measures of fit such as Akaike's and Schwarz's

Bayesian information criteria are not available. For the purpose of having a single measure of fit that could be applied to both the SFA models estimated using the *xtfrontier* command, and the LSE models, a pseudo-adjusted R^2 statistic was used to measure goodness-of-fit.[13]

Evoenergy did not respond to the consultation on the 2022 memo. The submission of AusNet emphasised the goodness-of-fit criterion, suggesting that:

> "the AER should consider removing the translog models in its entirely [sic], as it would free up resources to address more material concerns … The translog models do not improve on the fit of the Cobb-Douglas models" (AusNet, 2023).

It is possible for a group of additional explanatory variables to be added to a model, and be found to be statistically significant, and yet the goodness-of-fit statistic may not improve. This is because goodness-of-fit statistics often have a penalty for degrees of freedom lost. Adjusted R^2 statistics, and Akaike's and Schwarz's Bayesian information criteria mentioned above, all have such a penalty. Although a group of additional variables may be jointly statistically significant, the degree to which they improve the explanatory power of the model may not exceed the penalty in the goodness-of-fit statistic for degrees of freedom lost. Hence, these two criteria do not necessarily yield the same result.

Given these observations, the TLG specification is not unambiguously better than the CD specification. The added higher-order terms in the TLG model are in most cases jointly significantly different from zero, which indicates that there may be some nonlinearities that the linear-in-logs CD model does not capture. On the other hand, the CD model has the benefit of greater parsimony which enables it to rival the TLG model in terms of goodness-of-fit in most cases. And the CD model does not have the issue of monotonicity violations. This last advantage motivated Jemena, in its submission to the AER's 2022 draft annual benchmarking report, to propose that the Translog models should not be used in the benchmarking analysis (Jemena, 2022).

These considerations show there are different views on the relative performance of the CD and TLG models, which are based on their performance against different criteria. The TLG model has the advantage that joint significance tests usually suggest that the additional variables added to the model are significant. However, the CD model is more parsimonious and because of the loss of degrees of freedom in the TLG specification, it does not outperform the CD model in terms of goodness-of-fit. The CD model has the advantage that it is not found to suffer from monotonicity violations.

[13] The Quantonomics benchmarking study referred to by Evoenergy uses a Wald test of the restriction in the CD model that all the coefficients of the higher-order terms in the TLG model equal zero. There are two other commonly used methods, the Likelihood Ratio test and the Lagrange Multiplier test. The former can be applied using Stata's *lrtest* command while the latter can be applied using the contributed command *lgrgtest*, by Harold Tauchmann. However, these alternative testing methods cannot be applied after Stata's *xtpcse* command.

Finally, it is difficult to reconcile Evoenergy's argument that the TLG model is to be preferred over the CD model, based on the results of the Wald test of the joint significance of the higher-order terms, with its views expressed in relation to monotonicity violations. As discussed in section 3.1.3, Evoenergy asserts that the variation in the rates of monotonicity violations in the TLG models when applied to different benchmarking periods casts doubt on the reliability of all of the TLG opex cost function models, not just those with monotonicity violations. That is, Evoenergy suggests, the TLG models are to be rejected in their entirety because in some samples, for some DNSPs, they produce excessive monotonicity violations. How can this view be reconciled with the argument that the TLG models should be preferred to the CD specification because the higher-order terms in the TLG model are found to be jointly statistically significant?

In our view, at the present time, there is insufficient basis for entirely excluding the CD models or the TLG models, since they each have strengths and weaknesses when more than one criterion is used in evaluating them.

3.3 Stability of output weights and efficiency scores

The output weights used for both the base-step-trend opex forecasting method and the historical benchmarking roll-forward model are derived from the output elasticities of the opex cost function models. Evoenergy has three arguments relating to output weights:

- The output weights derived from the coefficients of the opex cost function models vary considerably between the four econometric models. Hence, if one or more of the models is excluded due to excess monotonicity violations, this can have a substantial effect on the base-step-trend and benchmarking roll-forward calculations.
- With the Translog models, the estimated output weights, when averaged by country, differ significantly between Australia, New Zealand and Ontario DNSPs and are strongly influenced by the latter. Hence, the sample average output weights "do not necessarily reflect well the output weights for Australian DNSPs" (Evoenergy, 2023a: 16).
- The differences between estimates obtained using the data sample 2006 to 2021 with those using the 2012 to 2021 sample suggests that "the output weights are not stable over time. Hence, the output weights calculated from the historical data may not be representative of the output weights that apply over an upcoming regulatory period" (Evoenergy, 2023a: 16).[14]

[14] Evoenergy says there is a 75 per cent overlap between the sample. However, since the sample from 2006 to 2021 has 16 observations and the sample from 2012 to 2021 has 10 observations, the overlap is 62.5 per cent. This does not have any bearing on Evoenergy's argument.

3.3.1 Effect on base-step-trend forecast of excluding output weights

Evoenergy is incorrect to suggest that when a model is excluded for the purpose of calculating efficiency scores, due to monotonicity violations, this can affect the base-step-trend calculations. In the base-step trend calculations, the AER uses average output weights from all four full sample models, without excluding models for monotonicity reasons. The same weights are used for all DNSPs.

In the historical benchmarking roll-forward model, the output weights used are DNSP-specific, and both the long-period and short-period models are used. However, models with excess monotonicity violations are removed from the calculation. The issue of effect on the historical roll-forward model of excluding models due to monotonicity violations has similarities to the issue of the effect on efficiency scores of excluding models which is discussed in section 3.1.2. Many of the same considerations discussed there, including the potential to substitute a constrained TLG model for an omitted TLG model, are relevant also here.

3.3.2 Variation of output weights between jurisdictions

In the base-step-trend forecast method, the output weights used by the AER are the average for the whole sample of DNSPs. Evoenergy argues that the output weights vary significantly between the countries in the sample, and hence the industry average weights are heavily influenced by the presence of those other countries in the sample, such as Ontario.

The estimated output elasticities averaged by jurisdiction are shown in Table 3.5 of the 2022 benchmarking study (Quantonomics, 2022a: 37). The output elasticities for each of the three outputs are used to calculate output weights by dividing each elasticity by the sum of the three output elasticities. This produces weights that sum to unity. Table 3.2 below shows the output weights from the four long-period models, for the sample overall, and for the Australian DNSPs.

Table 3.2 Average output weights calculated across groups of DNSPs, 2006-21

Output	*Full sample*	*Australian DNSPs*
Customer numbers	49.2%	43.6%
Circuit length	13.0%	17.4%
RMD	37.8%	39.0%
Total	100.0%	100.0%

Table 3.2 shows that there are some differences between the weights depending whether the full sample average is used or the average for the Australian DNSPs. The weight for customer numbers averaged over the Australian DNSPs is 5.6 percentage points lower than when averaged over the full sample. The weight for circuit length for the Australian DNSPs is 4.4

percentage points higher than for the full sample. The weight for RMD for Australian DNSPs is 1.2 percentage points higher than for the full sample.

With regard to which set of weights should be used in calculating the trend component of the base-step-trend forecast, Economic Insights has previously said "there is good economic justification for using the full sample mean Translog output weights in the rate of change calculations" (Economic Insights, 2020c: 18). However, it also added:

> "We are relatively indifferent as to whether the translog opex cost function output weights are calculated at the overall sample mean or at the Australian sample mean. We have demonstrated that there is economic justification for using either basis and the statistical performance of the models using either basis is little different. There may be some presentational and communication advantages in normalising by the Australian sample mean …" (Economic Insights, 2020c: 20).

We would reiterate the same views.

3.3.3 Variation of output weights over time

This section discusses Evoenergy's concern that the output weights for the models estimated over the shorter sample period (2012 to 2021) are substantially different from those estimated using the longer sample (2006 to 2021). Based on this observation, Evoenergy suggests that the output elasticities are changing over time and estimates derived from historical data may not be relevant for forecasting.

The opex cost function model is estimated using data for overseas DNSPs to ensure there is sufficient data to reliably estimate the underlying cost function (Economic Insights, 2020a: 9). By the same token, the long sample period, with considerably more observations, may be expected to provide a somewhat more reliable estimate of the cost function than using a shorter period. That said, the shorter sample period has one advantage compared to the long-period model. The estimated efficiency score for a DNSP is taken to represent its 'period average' efficiency over the sample period. It can take some time for more recent improvements in efficiency by previously poorer performing distribution businesses to be reflected in period-average efficiency scores. Hence the more recent time period may better capture more recent levels of operating efficiency. The AER uses both the long-period and short-period models for efficiency score estimates to balance the need for the most reliable estimates drawn from the largest available dataset, against the benefit of a more current estimate of efficiency that can be obtained from the shorter-period model, although estimated with slightly less precision.

These applications of the models for historical estimates of efficiency are quite different to applications of output weights for forecasting. Evoenergy assumes that the differences in the output elasticities obtained using the long and short sample periods are due to the output elasticities changing over time. However, it may be due to the elasticities estimated using the smaller sample having less precision. For the purpose of combining output forecasts into an output index, the AER uses the output weights derived from the longer sample period. This is a reasonable approach on account of:

- Simplicity, in light of there not being substantial differences in output weights between the long and short periods; and
- The longer sample output weights are expected to be estimated with more precision.

3.4 Degree of reliance on overseas data

Evoenergy has concerns that the majority of the observations in the dataset used for benchmarking are for Ontario DNSPs and only approximately 20 per cent of the observations relate to Australian DNSPs. Given the differences between Australian and overseas DNSPs, especially those in Ontario, it argued that this makes the results less representative of Australian DNSPs. This is a subject that has been discussed at length in past reports and memoranda by Economic Insights.

The reason for including international data in the opex cost function modelling is Economic Insights' (2014) finding that the limited time-series variability within the Australian data made it infeasible to reliably estimate the underlying cost function using only Australian data. International data are included not for the purpose of international benchmarking, but to add sufficient cross-sectional variations in the data to reliably estimate the individual parameters in the opex cost function modelling.

Economic Insights chose to include DNSPs from New Zealand and Ontario because there was existing published data for these DNSPs on a comparable basis to Australian DNSPs. Data for DNSPs in the USA was examined but found to be incomplete in key areas, such as circuit length. Furthermore, many network businesses in the USA are vertically integrated with generation or other activities and thus cost allocation issues may be significant. For these reasons, US DNSPs were not included. Economic Insights tested samples of Ontario and New Zealand DNSPs based on different size thresholds. For example, the medium dataset included overseas DNSPs with more than 20,000 customers. The small dataset included overseas DNSPs with more than 50,000 customers. The medium dataset was chosen.

Experience shows that the reliable estimation of econometric benchmarking models typically requires large data samples with sufficient data variations. The inclusion of overseas DNSPs has an important role in ensuring that there is a sufficient number of observations, and also sufficient diversity of characteristics among the DNSPs included, to improve the precision of the estimates of individual parameters of interest for Australian DNSPs.

It is appropriate to draw on an international sample of electricity network operators carrying out the same functions because they use common technology, which determines the shape of the opex cost function. Economic Insights stated:

> "we consider that the technologies used in distributing electricity across the three countries are common such that the output-cost relationship is not materially different. The inclusion of country dummy variables in the econometric models allows for systematic differences in operating environments between countries. Where operating conditions differ, this is likely to affect total opex in levels, rather than the output coefficients. An example of this would be Ontario's considerably harsher winter conditions which require more to be spent on clearing lines of ice and snow and keeping access to customers open. This would be likely to increase opex for an Ontario DNSP that was otherwise of similar size (or output mix) to an Australian DNSP or Ontario DNSPs as a group relative to Australian DNSPs as a group. However, for otherwise identical DNSPs, one in Australia and the other in Ontario, the same 1 per cent increase in line length, is expected to result in the same percentage increase in opex" (Economic Insights, 2020a: 9).

The commonality of output coefficients between jurisdictions implies, for example, that for otherwise identical DNSPs, the same 1 per cent increase in line length, is expected to result in the same percentage increase in opex. Economic Insights also explained that it is quite common for economic benchmarking models used in regulation to include a data sample of DNSPs from a diverse range of jurisdictions.

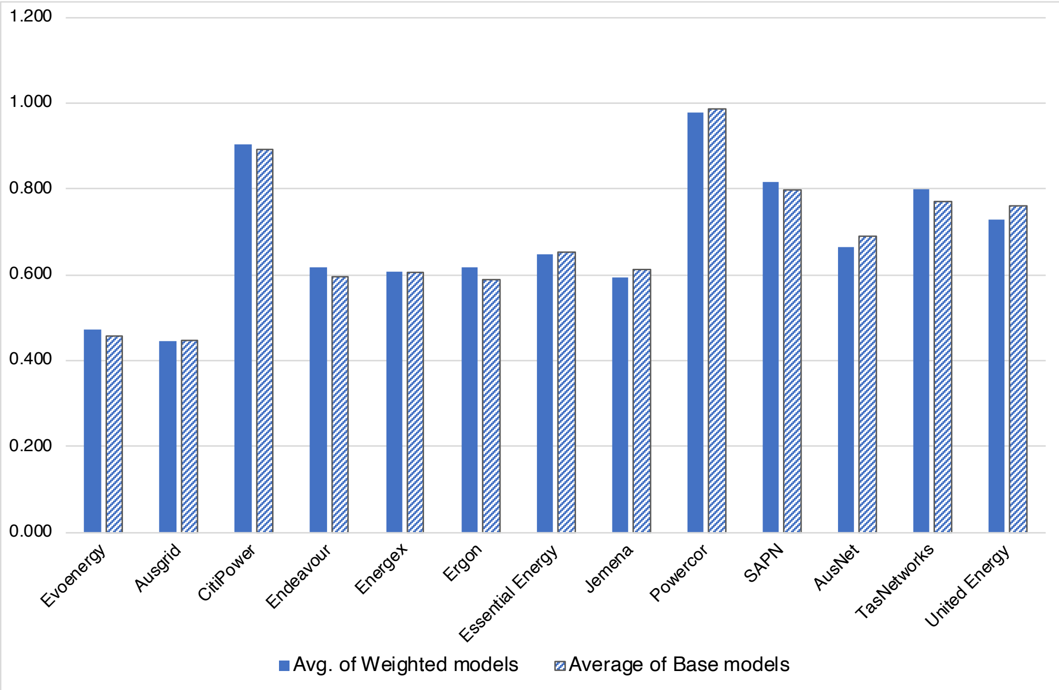
Evoenergy's claims that the underlying cost-output relationship is different for overseas DNSPs are not convincing due to a lack of supporting evidence. On the other hand, in previous exploratory work for the AER, Economic Insights (2021a) tested the effect of the overall weight given to Ontario DNSPs, although this research was not published. Economic Insights tested the approach of giving a larger weight to observations for Australian DNSPs in the regression analysis, than for the overseas DNSPs. The observations on Australian DNSPs were given a weight of 4, whereas for overseas DNSPs a weight of 1 was retained.[15] This resulted in Australian DNSPs having a weight of 48.9 per cent overall, or close to half the weight of the sample. In all other respects, the model specifications were unchanged and the CD and TLG models were estimated using both SFA and LSE estimation methods. This analysis found that while the weighted regression models did not solve the problem of monotonicity violations, the estimated efficiencies were similar to the results of the unweighted base model for all Australian DNSPs. This result is shown in Figure 3.1, which is reproduced from the Economic Insights memo.

It is possible to approach this issue in other ways, such as by excluding part of the overseas sample, for example, by adopting a higher size threshold; eg, only including DNSPs in New Zealand and Ontario with more than 50,000 customers. However, this would reduce the diversity of characteristics of DNSPs, which is important for reliably estimating the relationship between output and cost over a wide range of different output values. Further,

[15] This was implemented using Stata's option for incorporating 'analytic weights' (see: StataCorp, 2019 sec. 11.1.6).

simply by reducing the model degrees of freedom this may lead to less reliable parameter estimates.

Figure 3.1: Efficiency scores of Australian DNSPs in the weighted model (2006–2020)



Source: Economic Insights.

We cannot see a compelling reason for concluding that results using a smaller sample of DNSPs would be more reliable. Therefore, based on analysis undertaken to date, it is our view that the existing sample of Australian and international DNSPs remains appropriate at the present time.

The jurisdictional dummy variables currently used capture the combined effect of all *static* differences between jurisdictions not accounted for by other explanatory variables in the regression model. This includes effects of climate differences that are constant across the sample period. Whether there is merit in controlling for differences in underlying trends between jurisdictions is a matter for further investigation. Furthermore, it may be worth investigating, as part of the benchmarking development program, whether there is a suitable climate variable that can be identified and measured that can assist to further control for weather variations and jurisdictional differences from year-to-year.

In summary, Evoenergy's claims that the underlying cost-output relationship is different for overseas DNSPs are not convincing due to a lack of supporting evidence. On the other hand, there is evidence that when tested with a larger weight being applied to observations for Australian DNSPs in the regression analysis, so that the relative weight attributed to overseas DNSPs is reduced, there is only a small effect on estimated efficiency scores. This supports the conclusion that the importance of New Zealand and Ontarian DNSPs in the sample does not render the efficiency results unreliable for Australian DNSPs.

3.5 Input substitution

Evoenergy argues that the opex cost function model does not account for substitution between opex and capital inputs. As a result, a business that makes allocatively efficient choices to undertake more maintenance work rather than replacing assets will appear less efficient. Further, it considers these issues are not addressed by the AER's final decisions in its recent capitalisation review (AER 2023b).

In that review, the AER decided to address differences in capitalisation practices between DNSPs by allocating all corporate overheads expenditure to opex for benchmarking purposes. The AER noted that this change "does not comprehensively capture all sources of capitalisation differences, particularly in relation to opex/capital trade-offs. While opex/capital trade-offs are to some extent indirectly taken into account in our econometric opex cost function models, the extent to which opex/capital trade-offs are taken into account is unknown" (AER 2023b, iv).

The opex cost function developed by Economic Insights is in the nature of a short-run function, consistent with the building block model used in regulatory price reset decisions. In theory, a short-run input demand function for non-capital inputs will generally include as an explanatory variable the quasi-fixed quantity of capital input. However, Economic Insights chose not to include a capital input variable, partly because of statistical limitations due to a high correlation between the fixed input and the other outputs already included in the regression, and also the difficulty of obtaining data for comparable measures of capital input in each jurisdiction (especially Ontario, where assets are valued at depreciated historical cost). With outputs defined functionally, and with a substantial part of network outputs being the provision of delivery capacity over a given spatial dimension, there will invariably be a high correlation between outputs and capital input.

While Evoenergy observes that a business making an efficient decision to undertake more maintenance rather than replace an asset will appear less efficient, such decisions relating to substitution between inputs can also be made inefficiently. It is necessary to be able to distinguish between the effects of efficient and inefficient choices. This is an empirical question. Furthermore, the extent of substitutability between opex and capex is also an empirical question.

The AER has generally considered issues of substitution between opex and capex in the context of price resets, based on the information submitted to it. The substitution possibilities between operating and capital expenditure are one of the matters the AER is required to consider when deciding whether an operating expenditure forecast meets the relevant criteria (ie, being consistent with efficiency and prudency and based on realistic forecasts of demand

and input prices).[16] It need not be part of the benchmarking framework unless the degree of substitutability between capital and non-capital inputs is substantial.

The degree of substitution between capital and non-capital inputs may be able to be estimated within a long-run opex cost function model by incorporating the relative prices of capital and non-capital inputs. From the parameter on the relative price term, and depending on the functional specification, the elasticity of substitution can in principle be derived. There may be practical challenges in implementing such an approach, such as whether comparable data for input prices can be obtained for all three jurisdictions included in the data sample, and whether there is sufficient variation in relative input prices so as to permit the substitution effect to be reliably estimated. Until such an analysis is carried out, there is little information on the degree of substitutability between capital and non-capital inputs for electricity DNSPs.

Within a short-run opex cost function, a measure of capital stock is sometimes included, with opex being conditional on the 'quasi-fixed' capital input. The term 'quasi-fixed' is used because capital is not actually constant over the sample period, but it is treated as exogenous in the short-term. Over a long-sample period, the changes in the capital stock will be influenced by relative input prices, and hence their effect should be indirectly reflected in the short-run opex model, when it is estimated over a sufficiently long sample period. This approach does not provide information on the degree of substitution between capital and non-capital inputs, but substitution effects may to some extent be indirectly accounted for. As previously mentioned, there are practical difficulties in implementing this approach as a part of the AER's benchmarking models due to inconsistent measurement of capital between jurisdictions and the problem of multicollinearity between capital input and the output variables.

At this stage we have no information to suggest that substitutability between capital and non-capital inputs is substantial, and therefore no basis for concluding that the omission of long-run opex/capital substitution effects has a material effect on the reliability of opex cost function efficiency scores. However, it would be desirable to explore this question empirically by estimating a long-run opex cost function using a price ratio between capital and non-capital inputs. This is a substantial task that could form part of the benchmarking development program.

4 The use of benchmarking results in regulatory decisions

This section responds to the following views expressed by Evoenergy. "Whilst the AER's benchmarking analysis is informative and important, the AER should recognise that the benchmarking models suffer from many significant limitations" (Evoenergy, 2023a: 35). "To assess the efficiency of costs, base year revealed opex provides a reasonable estimate of the

[16] National Electricity Rules, clause 6.5.6(e)(7).

efficient and recurrent costs required to provide safe and reliable services while meeting regulatory obligations" (Evoenergy, 2023b: 20).

The AER's primary tool for establishing efficient expenditure is to rely on the incentive mechanisms within the regulatory framework to ensure DNSPs improve efficiency over time. If these incentives are sufficiently strong, the AER can rely on revealed costs as indicating efficient costs. Although this is the primary approach, it is supplemented by benchmarking and trend analysis to determine whether there is evidence that base year opex is materially inefficient and by imposing a general productivity improvement when forecasting opex.

The AER's current practice when examining whether opex in the base year is materially inefficient is to adjust the estimated efficiency scores by dividing them by an efficiency target of 0.75, which is then further adjusted for material OEFs. This means that, after accounting for OEF differences, only DNSPs with average efficiency scores of less than 0.75 are treated as inefficient and their efficiency scores are divided by 0.75 (ie, multiplied by 133.3 per cent) before they are applied.[17] This 'normalisation' procedure recognises that estimated efficiency scores are not exact and the true degree of relative efficiency is subject to some uncertainty. This approach should mitigate the effect of imprecision of efficiency estimates on the achievability of opex benchmarks.

In the AER's recent consultation on incentive mechanisms (AER 2023, 5), the AER found that the data showed incentive schemes had driven significant improvements in performance through efficiency gains, including opex being lower by 30 per cent per customer since 2011-12. In the review, consumer stakeholders questioned whether the incentives of networks to achieve economic efficiency are sufficiently strong, and specifically, whether the 75 per cent efficiency target should be increased. The AER considered this a question to be reviewed as part of the benchmarking development program (AER 2023, 5).

Ultimately, the uncertainty of efficiency assessments needs to be accounted for in some way, including by the use of an efficiency target that is substantially less than 1.0 (eg, the current 0.75). The AER's adjustment to the efficiency scores by dividing by 0.75 provides margin for errors in the efficiency estimates, and implies a high likelihood that the resulting adjusted scores are not underestimated.

These observations indicate that the AER *does* account for the limitations of benchmarking in its use of efficiency estimates. This finding is inconsistent with Evoenergy's claim that the AER does not adequately account for the limitations of the benchmarking analysis. In our view, the AER's procedure of setting the efficiency target at 0.75 adequately accounts for the potential errors of the estimated efficiency scores.

[17] For each inefficient DNSP (ie, having an efficiency score (ES) less than 0.75), the AER calculates: (a) the adjusted comparison point (ACP), equal to 0.75 adjusted for material OEFs; and (b) the efficiency target (ET) which is the ES divided by the ACP. Ignoring the OEF adjustment, ET = ES/0.75.

The question of whether revealed costs can alone provide a fully adequate basis for opex forecasting has been addressed by the AER in its review of the regulatory incentive mechanisms.

It is noted that the process of averaging the scores estimated by different models minimises model uncertainty. Uncertainty about the model itself is not amenable to statistical testing using the standard errors of the estimated parameters (Greene, 2012: 181), because these tests assume that the model specification is known. As discussed in section 3.2, the AER's approach of averaging models reduces the potential errors associated with model specification.

Appendix A: Outliers and Monotonicity Violations

In this appendix, the relative importance of outliers is first assessed, and then the relationship between monotonicity violations and outliers is examined.

Two alternative ways of identifying outliers are criteria based on the Interquartile Range (IQR) and on the size of Standardised Residuals. Both are shown in Table A.1 for four models, namely the LSETLG and SFATLG models in both the long and short sample periods. Criteria for mild and severe outliers based on the IQR are as defined in the user-contributed Stata command *iqr* by Lawrence Hamilton.[18] The criterion for outliers using standardised residuals is > 3 in absolute value, which is a commonly used criterion.

The statistics in Table A.1 show the distribution of the residuals has fatter tails than the normal distribution, which is common in econometric analysis of panel data. This implies a relatively high number of outliers. For example, 4.35 per cent of the residuals in the long-period SFATLG model are classified as mild outliers, which is much greater than the 0.7 per cent expected under a normal distribution.

Table A.1 Normality of residuals and outliers

	Sample period 2006–2021				*Sample period 2006–2021*			
	LSETLG model		*SFATLG model*		*LSETLG model*		*SFATLG model*	
	#	%	#	%	#	%	#	%
*IQR test**								
Mild outliers	30	2.80	33	3.07	13	1.95	29	4.35
Severe outliers	2	0.19	3	0.28	0	0.00	2	0.30
Standardised residuals								
>3.0 absolute value**	11	1.02	15	1.40	10	1.50	13	1.95

Notes: * Mild outliers make up about 0.7%, and severe outliers comprise about 0.0002%, of a normal distribution.
** Approximately 0.26% of a standard normal distribution will be outside this range.

[18] Denoting the 75th percentile as e(75) and the 25th percentile as e(25). IQR = e(75) – e(25). A mild outlier is: < e(25) – 1.5.IQR or > e(25) + 1.5.IQR. A severe outlier is: < e(25) – 3.IQR or > e(25) + 3.IQR.

In the following examination, outliers are defined as standardised residuals greater than 3.0 in absolute value. Table A.1 shows that this is wider than the definition of 'severe outliers' using the IQR, but narrower than the definition of 'mild outliers'.

Tables A.2 to A.5 show cross-tabulations of the number of monotonicity violations and the number of outliers for each of the same four models. Associated with each cross-tabulation is a Cramer's V statistic, which is a measure of association between binary or ordinal variables. It has a value of 0 if there is no association and 1 if there is perfect association. Values 0.1 to 0.3 are taken to represent weak association. Values 0.3 to 0.6 represent a moderate association, and values > 0.6 represent a strong association. The statistic takes positive or negative values depending on whether the association is positive or negative.

Table A.2 Outliers and monotonicity violations: LSETLG 2006–2021

	Outliers		
Monotonicity violations	*No*	*Yes*	*Total*
No	937	3	940
Yes	126	8	134
Total	1,063	11	1,074

Table A.3 Outliers and monotonicity violations: SFATLG 2006–2021

	Outliers		
Monotonicity violations	*No*	*Yes*	*Total*
No	929	15	944
Yes	130	0	130
Total	1,059	15	1,074

Table A.4 Outliers and monotonicity violations: LSETLG 2012–2021

	Outliers		
Monotonicity violations	*No*	*Yes*	*Total*
No	530	2	532
Yes	126	8	134
Total	656	10	666

Table A.5 Outliers and monotonicity violations: SFATLG 2012–2021

	Outliers		
Monotonicity violations	*No*	*Yes*	*Total*
No	608	13	621
Yes	45	0	45
Total	653	13	666

The Cramer's V statistics associated with these cross-tabulations are:

- LSETLG 2006–2021: 0.1855

- SFATLG 2006–2021: –0.0442
- LSETLG 2012–2021: 0.1844
- SFATLG 2012–2021: –0.0380

These results show that the measures of association differ between the LSETLG and SFATLG models but are consistent between the long and short sample periods. In the LSETLG models, the Cramer's V statistic is less than 0.2, indicating only a very weak association between outliers and monotonicity violations. In the SFATLG models, the Cramer's V statistic is close to zero, indicating no association between outliers and monotonicity violations.

References

Agrell P and Bogetoft P (2009) *International Benchmarking of Electricity Transmission System Operators: e3Grid Project Final Report*. March. Sumicsid.

Agrell PJ and Bogetoft P (2016) Regulatory Benchmarking: Models, Analyses and Applications.

AusNet (2023) *Submission to the AERs memorandum on technical issues with the econometric opex cost models*. 9 February.

Australian Competition Tribunal (2016) *Applications by Public Interest Advocacy Centre Ltd and Ausgrid [2016] ACompT 1.*

Australian Energy Regulator (AER) (2015) *Ergon Energy determination 2015-16 to 2019-20: Preliminary decision (Attachment 7, Operating expenditure)*. April.

Australian Energy Regulator (AER) (2021) *Jemena Distribution Determination 2021-22 to 2025-26: Final decision (Attachment 6, Operating expenditure)*. April.

Australian Energy Regulator (AER) (2022) *Annual Benchmarking Report: Electricity distribution network service providers*. November.

Australian Energy Regulator (AER) (2023a) *How the AER will assess the impact of capitalisation differences on our benchmarking: Final Guidance note*. May.

Australian Energy Regulator (AER) (2023b) *Review of incentives schemes for networks: Final decision*. April.

Azadeh A, Ghaderi SF, Omrani H, et al. (2009) An integrated DEA–COLS–SFA algorithm for optimization and policy making of electricity distribution units. *Energy Policy* 37(7): 2605–2618.

Cadena A, Marcucci A, Perez JF, et al. (2009) Efficiency Analysis in Electricity Transmission Utilities. *Journal of Industrial and Management Optimization* 5(2): 253–274.

Chatterjee S and Hadi AS (1986) Influential Observations, High Leverage Points, and Outliers in Linear Regression. *Statistical Science* 1(3): 379–393.

Diewert E (2009) The aggregation of capital over vintages in a model of embodied technical progress. *Journal of Productivity Analysis* 32(1): 1–19.

Economic Insights (2013a) *Measurement of Outputs and Operating Environment Factors for Economic Benchmarking of Electricity Distribution Network Service Providers*. 16 April. Briefing Notes prepared by Denis Lawrence and John Kain for Australian Energy Regulator.

Economic Insights (2013b) *Outputs and Operating Environment Factors to be Used in the Economic Benchmarking of Electricity Distribution Network Service Providers*. Briefing Notes prepared by Denis Lawrence and John Kain for Australian Energy Regulator, 20 February.

Economic Insights (2014) *Economic Benchmarking Assessment of Operating Expenditure for NSW and ACT Electricity DNSPs*. Report prepared for Australian Energy Regulator by Denis Lawrence, Tim Coelli and John Kain, 17 November.

Economic Insights (2015) *Response to Consultants' Reports on Economic Benchmarking of Electricity DNSPs*. Report prepared by Denis Lawrence, Tim Coelli and John Kain for the Australian Energy Regulator, 22 April.

Economic Insights (2017) *Topics in efficiency benchmarking of energy networks: Selecting cost drivers*. Report prepared by Denis Lawrence, John Fallon, Michael Cunningham, Valentin Zelenyuk, Joseph Hirschberg for The Netherlands Authority for Consumers and Markets, December.

Economic Insights (2020a) Comments on 2019 Frontier Economics Benchmarking Reports for EQ.

Economic Insights (2020b) *Economic Benchmarking Results for the Australian Energy Regulator's 2020 DNSP Annual Benchmarking Report*. Prepared for Australian Energy Regulator by Denis Lawrence, Tim Coelli and John Kain, October.

Economic Insights (2020c) *Review of reports submitted by CitiPower, Powercor and United Energy on opex input price and output weights*. 18 May.

Economic Insights (2021a) *DNSP opex cost function: Exploring alternative specifications*. Prepared for the AER by Michael Cunningham, 29 September.

Economic Insights (2021b) *Economic Benchmarking Results for the Australian Energy Regulator's 2021 DNSP Annual Benchmarking Report*. Draft report prepared by Michael Cunningham, Denis Lawrence and Tim Coelli for Australian Energy Regulator, 12 November.

Evoenergy (2022) Submission on the impact of capitalisation on the AER's benchmarking.

Evoenergy (2023a) *Regulatory proposal for the ACT electricity distribution network 2024–29, Appendix 2.1: Operating expenditure base year efficiency*. 31 January.

Evoenergy (2023b) *Regulatory proposal for the ACT electricity distribution network 2024–29, Attachment 2: Operating expenditure*. January.

Evoenergy (2023c) Submission on the impact of capitalisation on the AER's benchmarking.

Farsi M, Fetz A and Filippini M (2007) *Benchmarking and regulation in the electricity distribution sector*. 54, CEPE Working Paper, January. Centre for Energy Policy and Economics, Swiss Federal Institutes of Technology.

Frontier Economics (2019) *AER Operating Environment Factors (OEFs)*. January. Prepared for Energy Queensland.

Greene WH (2012) *Econometric Analysis*. 7th ed. Pearson.

Haney AB and Pollitt MG (2012) *International Benchmarking of Electricity Transmission by Regulators: Theory and Practice*. EPRG 1226, Working Paper, November. University of Cambridge, Electricity Policy Research Group.

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

Jemena (2022) Submission to 2022 AER draft distribution benchmarking report. Available at: https://www.aer.gov.au/system/files/Jemena%20-%20Submission%20to%202022%20AER%20draft%20distribution%20benchmarking%20report%20-%2026%20October%202022.pdf.

Kumbhakar SC, Wang H-J and Horncastle AP (2015) *A Practitioner's Guide to Stochastic Frontier Analysis Using Stata*. Cambridge University Press.

Kuosmanen T (2012) Stochastic semi-nonparametric frontier estimation of electricity distribution networks: Application of the StoNED method in the Finnish regulatory model. *Energy Economics* 34: 2189.

Lowry MN and Getachew L (2009) Econometric TFP targets, incentive regulation and the Ontario gas distribution industry. *Review of Network Economics* 8(4).

Ofgem (2022) *RIIO-ED2 Draft Determinations – Core Methodology Document*. 29 June.

Orea L and Jamasb T (2017) Regulating Heterogeneous Utilities: A New Latent Class Approach with Application to the Norwegian Electricity Distribution Networks. *The Energy Journal* 38(4).

Oxera (2020) *A critical assessment of TCB18 electricity*. Prepared for the European electricity TSOs, 30 April. Available at: https://www.oxera.com/wp-content/uploads/2020/06/A-critical-assessment-of-TCB18-electricity_Oxera_Final.pdf.

Quantonomics (2022a) *Economic Benchmarking Results for the Australian Energy Regulator's 2022 DNSP Annual Benchmarking Report*. Report prepared for Australian Energy Regulator by Michael Cunningham, Joseph Hirschberg and Melusine Quack, 17 November.

Quantonomics (2022b) *Opex Cost Function Development*. Memo prepared for AER by Michael Cunningham and Joe Hirschberg, 7 October.

Sapere research group and Merz Consulting (2018) *Independent review of Operating Environment Factors used to adjust efficient operating expenditure for economic benchmarking*. Report for the Australian Energy Regulatory by Simon Orme, Dr. James Swansson, Geoff Glazier, Ben Kearney, Dr. Howard Zhang, August.

StataCorp (2019) *Stata User's Guide: Release 16*. Stata Press.

Sumicsid and Council of European Economic Regulators (CEER) (2019) *CEER-TCB18: Pan-European cost-efficiency benchmark for electricity transmission system operators*. July.

Yu W, Jamasb T and Pollitt M (2009) Does weather explain cost and quality performance? An analysis of UK electricity distribution companies. *Energy Policy* 37(11): 4177–4188.