Memorandum

From:	Tim Coelli and Denis Lawrence	**Date:**	11 March 2020
To:	AER Opex Team		
Subject:	Comments on 2019 Frontier Economics Benchmarking Reports for EQ		

Economic Insights has been asked to provide advice to the AER regarding two recent reports prepared by Frontier Economics (FE 2019a,b) for Energy Queensland (EQ). These reports were submitted to the AER by EQ in its initial and revised proposals, respectively, for the 2020–25 regulatory period.

FE (2019a,b) make a number of critiques of and suggestions regarding the AER's economic benchmarking of electricity distribution network service providers (DNDPs). The critiques and suggestions cover econometric and modelling issues and the operating environment factors used by the AER (OEFs). This assessment of the FE (2019a,b) reports covers only econometric and modelling issues. Specific OEF issues are being addressed separately by the AER.

We proceed to assess each of the main econometric and modelling issues raised by FE (2019a,b).

Tribunal recommendations

FE (2019b, p8) state that:

> *Nearly four years has passed since the Tribunal handed down its judgement. Yet, the benchmarking approach applied by the AER in its Draft Decisions fails to address many of the fundamental failings identified by the Tribunal. Many of those criticisms made by the Tribunal have received only cursory consideration by the AER or have been ignored altogether.*

We have two general comments to make here. First, the AER has made substantive changes in response to comments made by the Tribunal and during the Tribunal proceedings. For example:

- The AER now uses an average of predictions from up to four econometric models and, in some cases, total factor productivity index number models as well as other supporting tools including PPIs at total cost level and at cost category level (eg vegetation management, emergency response, maintenance, etc) to assess DNSPs' opex efficiency and potential opex productivity growth instead of just the one econometric model as criticised by the Tribunal
- The AER has undertaken a major review of the OEFs it uses and now focuses on key material OEFs instead of also including an allowance for immaterial OEFs, and

- The AER has moved to further improve data quality and consistency among DNSPs as the EBRIN data – which were of relatively high quality to start off with – are progressively refined as remaining minor issues come to light.

Second, we observe that the predictions from the SFACD model that were produced in Economic Insights (2014, 2015) and subsequently used by AER (2015) in forming its forecast of efficient opex for the NSW and ACT DNSPs have proven to be quite accurate. Although the Tribunal found the AER's claimed sole reliance of the SFACD model to be incorrect, the passage of time has shown that these regulated firms have actually achieved opex levels very similar to the original AER targets derived from the Economic Insights benchmarking models. In remaking the decisions in 2018, the AER observed that the NSW/ACT DNSPs have appeared to respond to the stronger incentives imposed by the use of economic benchmarking to reduce their opex to the level set by the AER's 2015 decision.

Bottom–up benchmarking

FE (2019b, p9) state that:

> *We note that the Tribunal directed the AER explicitly to remake its 2015 opex decisions (for those DNSPs that sought merits reviews) by undertaking a bottom–up review of forecast opex. The AER has not undertaken any bottom–up assessment of base year opex in the Draft Decisions, even as a cross–check of its top–down benchmarking analysis.*

The first thing to note here is that exactly what is meant by 'bottom–up' assessment can vary widely. At one end of the spectrum it can mean detailed engineering assessments at a relatively micro level of all the DNSP's processes. At the other end of the spectrum it can mean a much higher–level assessment of a number of categories of opex using techniques such as regression analysis, ratio analysis and/or unit cost assessment with multiple or single drivers identified as the key cost drivers at the activity level. The Tribunal did not define exactly what it meant by 'bottom–up' analysis.

During the Tribunal process the AER noted there had been limited use of bottom-up engineering modelling in earlier resets but this was generally applied to capex assessment while opex relied on the revealed cost approach to use past expenditure as the starting point. The DNSPs' submissions argued for a line–by–line, bottom–up review at the category level. As noted above, the AER has now moved to include PPIs at the total cost level and at the cost category level (eg vegetation management, emergency response, maintenance, etc) as supporting tools in its economic benchmarking.

For the purposes of this discussion we define 'bottom–up' analysis to include detailed engineering assessments at the process level. But similar issues will also apply to the somewhat higher–level analysis of individual opex categories.

A bottom–up assessment would likely involve a significant and onerous data collection activity for each regulated DNSP, which can often be costly and invasive. This type of process–level regulatory benchmarking can result in the formation of a second layer of administration within each DNSP, with the regulator essentially "second guessing" every decision made regarding the various individual activities within the DNSP.

In addition to this, it should be noted that a bottom–up performance assessment will generally involve the production of a range of technical indicators and partial productivity measures. Any assessment of a list of individual partial productivity benchmarking measures or technical indicators will in general produce an aggregate benchmarking target for a firm that is more stringent than that derived from the use of a global or "top–down" productivity measure, such as those calculated using frontier analysis.[1] This is because the individual assessment of a series of technical indicators or PPIs at the process level is likely to lead to 'cherry–picking' of results which risk producing an overall target that is unobtainable. This contrasts to 'tops–down' analysis where much better account is taken of trade–offs and feasibility constraints.

This can be shown using the simple illustration provided below in Table 1 and Figure 1. In the first section of Table 1 we have constructed artificial data for four DNSPs, labelled A, B, C and D. We assume that each DNSP uses only one input variable (10 units of labour) and two output variables (customer numbers and line length).[2] If a frontier analysis is conducted, Figure 1 indicates that DNSPs A, C and D lie on the estimated frontier and hence are deemed to be efficient, while DNSP B is inside the frontier with an efficiency measure of approximately 67 per cent, which is the degree by which it could proportionally expand its outputs (along the dotted line) or reduce its labour input and still remain within the production frontier.

If we instead assess the performance of these four DNSPs using the separate partial productivity ratios of customers/labour and lines/labour reported in the first section of Table 1 we conclude that only DNSP A is efficient in terms of lines/labour and only DNSP D is efficient in terms of customers/labour and also observe that no DNSP is efficient in both indicators – with efficient production now defined by the star marked on Figure 1. However, the 'tops–down' frontier analysis indicates that the starred point derived from the separate analyses of 'bottom–up' indicators is likely to be, at best, overly onerous and, at worst, infeasible.

Next, consider an alternative use of these 'bottom–up' partial productivity ratios where the DNSPs first attempt to allocate their total labour input across these two output activities prior to constructing the individual partial productivity ratios. This is considered in the lower part of Table 1 where we investigate two simple cases of labour allocation. In the first we assume each DNSP nominally allocates one employee per 100kms of line (with the remainder allocated to customer activities) while in the second we assume each DNSP allocates one employee per 25 customers (with the remainder allocated to lines activities). In each of these two cases we find that only one DNSP is efficient and the other three are inefficient. Compare this with the 'tops–down' frontier method where three out of the four DNSPs were observed to be efficient.

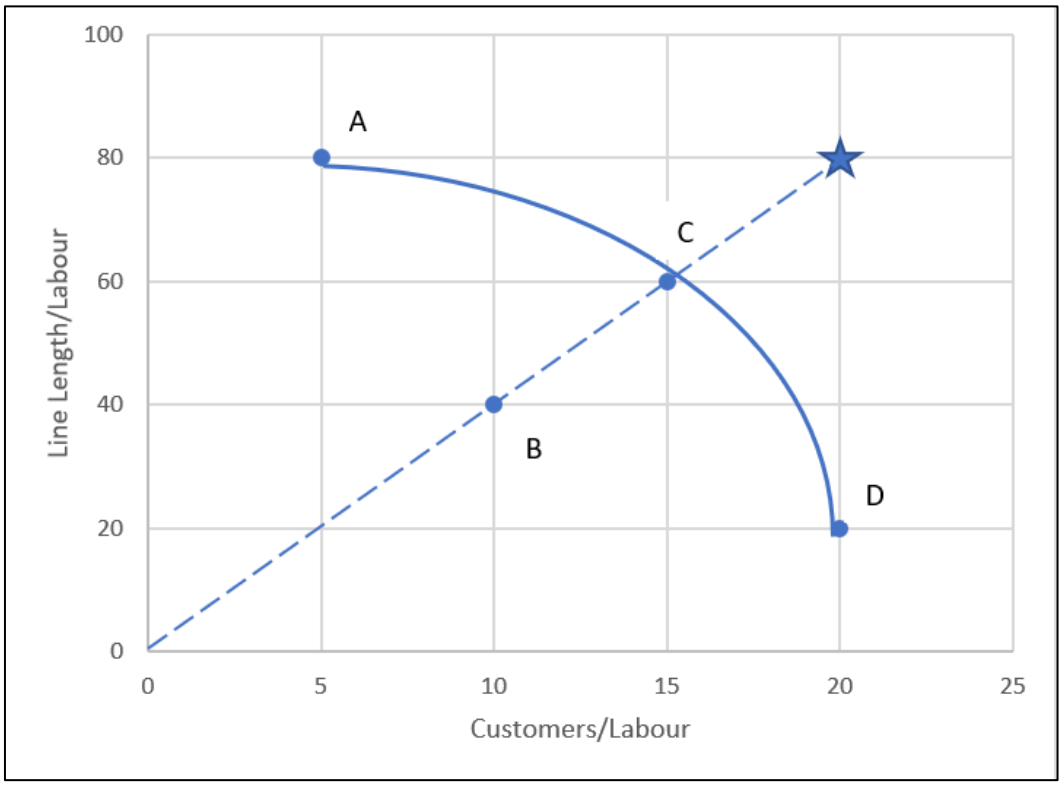
When input allocations are made across output categories like this another issue needs to be noted. That is, given that each DNSP generally decides on its own internal allocation rules, a DNSP might decide to utilise an unusual allocation method that allocates an unrealistically small amount of input to one output activity (eg lines) and a correspondingly larger amount to another output activity (eg customers). As a result, the partial productivity measure for lines might then define an efficiency level that is not technically feasible – and hence even further overstate the degree of inefficiency across the remainder of the group of DNSPs.

[1] This assumes the basis of assessment is efficient unit costs and the output specification is demand-driven.
[2] This allows us to draw this example in two dimensions.

Table 1 **Bottom–up versus top–down benchmarking**

DNSP	labour	lab cust	lab line	customers	line length	cust/lab	line/lab
A	10			50	800	5	80
B	10			100	400	10	40
C	10			150	600	15	60
D	10			200	200	20	20
DNSP	**labour**	**lab cust**	**lab line**	**customers**	**line length**	**cust/lab**	**line/lab**
A		2	8	50	800	25	100
B		6	4	100	400	17	100
C		4	6	150	600	38	100
D		8	2	200	200	25	100
DNSP	**labour**	**lab cust**	**lab line**	**customers**	**line length**	**cust/lab**	**line/lab**
A		2	8	50	800	25	100
B		4	6	100	400	25	67
C		6	4	150	600	25	150
D		8	2	200	200	25	100

Figure 1 **Bottom–up versus top–down benchmarking**



Given the above discussion, we would expect that it would normally be in the interests of a regulated DNSP – and also consumers – to argue for the use of aggregate 'tops–down' frontier

methods over the use of 'bottom–up' analyses of technical indicators and partial productivity measures. Hence, it is somewhat surprising to see FE (2019b) apparently arguing for the use of the latter approach.

Urban versus rural

FE (2019b, p10) state that:

> *We demonstrated in our January 2019 benchmarking report that statistical testing suggests that rural and urban samples should not be pooled together. This result is intuitively and economically compelling since rural and urban DNSPs typically operate in quite different environments, often have differently–engineered networks and face different cost challenges.*

We have studied the analysis referred to in the FE (2019a) benchmarking report and note that their claimed statistical test has been conducted only for the SFACD model using the 2006–2017 data.

There are a number of issues regarding this "poolability" analysis. First, FE (2019a,b) do not state what type of hypothesis test has been conducted. It is not indicated whether a Likelihood Ratio test or a Wald test or another asymptotic test with a chi–square distribution has been undertaken and we can find no record of this test in FE's accompanying Stata files.

Second, while the FE (2019b) report argues for the use of the Translog (TL) functional form and also for the use of the shorter data set 2012–2018, it does not provide any explicit updated urban-rural "poolability" tests for the data sets and models used in this more recent report. It only refers to a single test of the SFACD model on the longer 2006–2017 data set from the FE (2019a) report. It is not clear why the tests were not repeated and reported for the two time periods of 2006–2018 and 2012–2018 and for the four models of SFA CD, LSE CD, SFA TL and LSE TL. It is also not clear whether these tests provided the same answer and had stable parameter estimates.

Third, there is a lack of information on whether estimated first–order coefficients have the expected size and correct sign and whether monotonicity conditions are satisfied in the TL models. FE (2019a) do present the estimated coefficients of the SFACD model in their Appendix C, which does have correctly signed coefficients. However, the FE (2019b) report document presents no additional estimated coefficients for the other models nor do they provide monotonicity information for the TL models.

The estimated coefficients for these extra models were however available in the Stata files that were attached to the FE (2019b) report document.[3] These econometric estimates have been extracted and are summarised in Table 2 below. The first thing we note is that six of the eight models in this table have incorrectly signed estimated coefficients for ShareUGC. That is, these models imply that an increase in undergrounding results in an <u>increase</u> in opex. This is counter-intuitive and undermines the validity of these models. Furthermore, we note that the estimated first-order coefficients on RMDemand are also incorrectly signed in four Urban

[3] We note that the Stata files initially submitted by FE were written in a non–transparent format that also suppressed much of the information required to assess the FE analyses. The AER subsequently requested FE to resubmit the Stata files in a more transparent and unsuppressed format.

interaction models, suggesting that (at the sample mean) an increase in RMDemand will result in a reduction in opex. Again, this is counter-intuitive and undermines the validity of these models.[4]

Table 2 **Estimated first-order coefficients in baseline and urban-rural models**

Model	Baseline							
Period	2006-2018				2012-2018			
Method	CDLSE	TLLSE	CDSFA	TLSFA	CDLSE	TLLSE	CDSFA	TLSFA
CustNum	0.682	0.512	0.665	0.673	0.684	0.443	0.660	0.587
CircLen	0.154	0.152	0.149	0.144	0.179	0.195	0.217	0.202
RMDemand	0.153	0.303	0.173	0.152	0.134	0.332	0.107	0.183
ShareUG	-0.156	-0.145	-0.134	-0.103	-0.161	-0.129	-0.084	-0.054

Model	Rural							
Period	2006-2018				2012-2018			
Method	CDLSE	TLLSE	CDSFA	TLSFA	CDLSE	TLLSE	CDSFA	TLSFA
CustNum	0.553	0.546	0.837	0.626	0.504	0.309	0.550	0.271
CircLen	0.077	0.137	0.159	0.233	0.055	0.120	0.120	0.609
RMDemand	0.380	0.256	0.035	0.203	0.467	0.524	0.365	0.169
ShareUG	-0.218	-0.182	-0.150	-0.133	-0.231	-0.188	-0.219	-0.241

Model	Urban							
Period	2006-2018				2012-2018			
Method	CDLSE	TLLSE	CDSFA	TLSFA	CDLSE	TLLSE	CDSFA	TLSFA
CustNum	0.572	0.296	0.566	0.937	0.402	0.225	0.347	0.781
CircLen	0.482	0.400	0.322	0.176	0.722	0.753	0.709	-0.067
RMDemand	-0.033	0.357	0.041	-0.250	-0.087	0.086	-0.080	0.237
ShareUG	0.003	0.037	-0.178	-0.131	0.122	0.251	0.090	0.160

The signs of the first-order coefficients in the TL models provide information on the monotonicity conditions at the sample mean. However, monotonicity conditions can vary across observations in TL models. Hence a thorough monotonicity test requires the evaluation of elasticities at each and every data point in the sample. We therefore looked at the FE (2019b) Stata files to see if we could find information on the observation specific elasticities but were unable to find this information. Consequently, we have conducted these calculations ourselves and have summarised the results for the Australian DNSPs in Table 3 below. We observe that

[4] We also note that the coefficient of CircLen is incorrectly signed in the TLSFA 2012-2018 Urban model as well.

all of the Rural/Urban models have a substantial number of monotonicity violations, with very large proportions of observations of 33 per cent, 33 per cent, 36 per cent and 100 per cent in some cases, indicating significant problems with all of these models.

Table 3 **Monotonicity violations in Australian data in baseline and urban-rural Translog models, percentage of observations**

	2006-2018					
Method:	**LSE**			**SFA**		
Model	**CustNum**	**CircLen**	**RMDemand**	**CustNum**	**CircLen**	**RMDemand**
Base	0%	0%	0%	0%	1%	2%
Rural	0%	0%	0%	0%	0%	36%
Urban	0%	0%	0%	4%	0%	33%

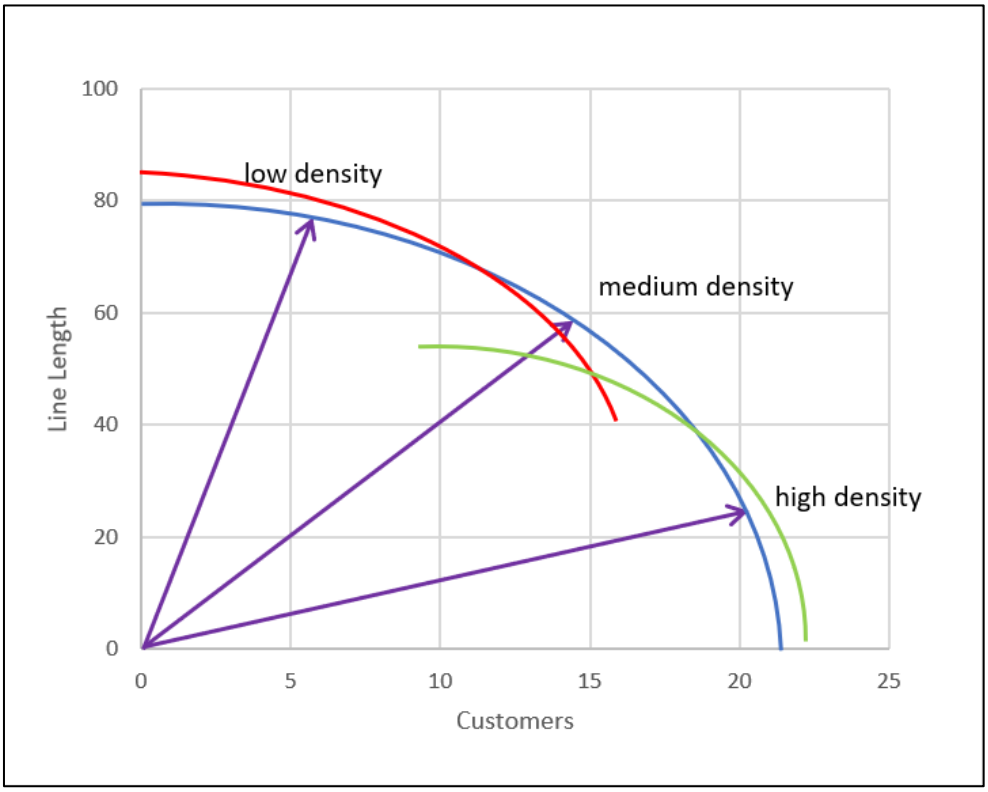
	2012-2018					
Method:	**LSE**			**SFA**		
Model	**CustNum**	**CircLen**	**RMDemand**	**CustNum**	**CircLen**	**RMDemand**
Base	25%	0%	0%	0%	3%	0%
Rural	0%	0%	0%	14%	14%	0%
Urban	100%	0%	0%	19%	0%	33%

Fourth, the pooling test that is reported appears to hinge on an arbitrarily selected point at which urban becomes rural, namely 20 customers per kilometre of circuit length. FE (2019a,b) provide no information on how sensitive the claimed test result is to adjustments in this arbitrary cut–off point. Also, a number of Australian DNSPs have a network that is a mixture of highly urbanised areas and rural areas. For example, the Tasmanian DNSP covers Hobart and Launceston as well as rural areas in the rest of the state. Similarly, the South Australian DNSP covers Adelaide – a city of well over one million – plus the rest of the state. And the two 'rural' Victorian DNSPs each cover significant sections of suburban Melbourne. Having an arbitrary cut–off of 20 customers per kilometre overall that classes these DNSPs as rural is unlikely to accurately the reflect the characteristics of these DNSPs that cover very different types of service areas and is likely to distort the model results. Furthermore, it is unclear whether the application of this arbitrary cut-off point is appropriate for the overseas DNSPs with only two NZ DNSPs being classified as urban and only two Ontario DNSPs being classified as rural.

As an illustration of the potential problems of dividing the sample into two groups, consider Figure 2 below. The frontier methods used in the AER work (for example illustrated by the blue curve in Figure 2) are designed to allow DNSPs with similar output mixes (eg customers per km of line) to be benchmarked with each other – providing a continuum of benchmarking subgroups as you move around the non–linear production surface. When the sample is divided

into artificial sub–groups and individual sub–group frontiers estimated (illustrated by the red and green lines in Figure 2), there is often little to be gained (even though the estimated parameters might differ to some extent). But some additional uncertainty can be introduced for those medium density firms that are located on the boundary of the two groups.

Figure 2 **Customer density across the frontier**



Fifth, it should be noted that by including both customer numbers and line length as outputs, we are allowing for differences in customer density across DNSPs. This point was actually well demonstrated in FE (2015, p.39) in an earlier report prepared for Ergon Energy. Because the functional forms we use are logarithmic, including customer numbers and line length as separate outputs is equivalent to including line length as an output along with customer density as an OEF. In effect, this is similar treatment of customer density to our treatment of the share of undergrounding OEF included in our models. This greatly reduces the need to consider separate treatment of rural and urban DNSPs, particularly given the potentially arbitrary and inadequate nature of the definition of rural versus urban.

Sixth, the accuracy of estimation is actually improved by having diverse characteristics in the sample. If all included DNSPs have similar characteristics then the model will find it hard to provide robust parameter estimates. This was noted in the current context by our fellow economic benchmarking practitioner, Pacific Economics Group (PEG 2008, p.12):

> *Notice also that the precision of an econometric benchmarking exercise is enhanced by using data from companies with diverse operating conditions. For example, we will obtain a better estimate of the impact of line length on cost if we include in the sample*

> *companies that, like Toronto Hydro Electric System (THES), have high customer density as well as data for companies that, like Sioux Lookout, have low density.*

In other words, instead of increasing the accuracy of treatment for rural DNSPs, attempts to separate samples for rural and urban DNSPs are likely to lead to less precise parameter estimates for all DNSPs.

This leads to our seventh point that it is quite common for regulatory economic benchmarking models to include both rural and urban DNSPs in the sample. Allowance may be made for differences in customer density across DNSPs either by choice of the output specification or by inclusion of an OEF variable. But it is rare for economic benchmarking studies to attempt to provide completely separate parameter estimates for most variables for rural and urban DNSPs. For example, the long line of economic benchmarking studies PEG has prepared for the Ontario Energy Board, culminating in PEG (2019a) include a sample of rural and urban DNSPs with a wide range of customer densities. Separate customer numbers and line length outputs are included but no allowance is made for differentiation of parameter estimates across rural and urban DNSPs.

Ontario data

FE (2019b, p10) state that they have:

> *shown on a number of occasions, including in our January 2019 benchmarking report, that statistical testing demonstrates that data on Ontarian DNSPs should not be pooled with data on Australian and New Zealand DNSPs.*

However, the application of 'poolability' tests in this instance is completely flawed. As explained in Economic Insights (2014), the reason for including international data in the opex cost function modelling is the inability to reliably estimate the underlying cost function using only Australian data, due to the limited time–series variability within the Australian data. Therefore, it is no surprise that the estimated coefficients from the Australian only or Australian and NZ models appear to be different from the full model in a 'poolability' test. But this is because the former cannot be reliably estimated.

Furthermore, 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.

Despite the inherently circular nature of undertaking a poolability test in this instance, we have studied the analysis referred to in the FE (2019a) benchmarking report and note that their

claimed statistical test has been conducted only for the SFACD model using the 2006–2017 data.

There are again several issues regarding this "poolability" analysis. First, FE (2019a,b) do not state what type of hypothesis test has been conducted. It is not indicated whether a Likelihood Ratio test or a Wald test or another asymptotic test with a chi–square distribution has been undertaken and we can find no record of this test in FE's accompanying Stata files.

Second, while the FE (2019b) report argues for the use of the Translog (TL) functional form and also for the use of the shorter data set 2012–2018, it does not provide any explicit updated "poolability" tests for the data sets and models used in this more recent report. It only refers to a single test of the SFACD model on the longer 2006–2017 data set from the FE (2019a) report. It is not clear why the tests were not repeated and reported for the two time periods of 2006–2018 and 2012–2018 and for the four models of SFA CD, LSE CD, SFA TL and LSE TL. It is also not clear whether these tests provided the same answer and had stable parameter estimates.

Third, there is a lack of information on whether estimated first-order coefficients have the expected size and correct sign and whether monotonicity conditions are satisfied in the TL models. FE (2019a) do present the estimated coefficients of the SFACD model in their Appendix C, which does have correctly signed coefficients. However, the FE (2019b) report presents no additional estimated coefficients for the other models nor do they provide monotonicity information for the TL models.

Table 4 **Estimated first-order coefficients in baseline and Ontario models**

Model	Baseline							
Period	2006-2018				2012-2018			
Method	CDLSE	TLLSE	CDSFA	TLSFA	CDLSE	TLLSE	CDSFA	TLSFA
CustNum	0.682	0.512	0.665	0.673	0.684	0.443	0.660	0.587
CircLen	0.154	0.152	0.149	0.144	0.179	0.195	0.217	0.202
RMDemand	0.153	0.303	0.173	0.152	0.134	0.332	0.107	0.183
ShareUG	-0.156	-0.145	-0.134	-0.103	-0.161	-0.129	-0.084	-0.054

Model	Ontario							
Period	2006-2018				2012-2018			
Method	CDLSE	TLLSE	CDSFA	TLSFA	CDLSE	TLLSE	CDSFA	TLSFA
CustNum	0.390	0.034	0.226	0.100	0.162	-0.202	0.189	-0.091
CircLen	0.494	0.375	0.261	0.247	0.720	0.593	0.495	0.408
RMDemand	0.046	0.546	0.421	0.540	0.046	0.559	0.209	0.640
ShareUG	0.150	0.183	-0.045	-0.038	0.295	0.394	0.192	0.226

The estimated coefficients for these extra models were however available in the Stata files that were attached to the FE (2019b) report. These econometric estimates have been extracted and are summarised in Table 4. The first thing we note is that six of the eight models in this table have incorrectly signed estimated coefficients for ShareUGC. That is, these models imply that

an increase in undergrounding results in an increase in opex. This is counter-intuitive and undermines the validity of these models. Furthermore, we note that the estimated first-order coefficients on CustNum are also incorrectly signed in two models (TLLSE and TLSFA using 2012-2018 data), suggesting that (at the sample mean) an increase in customers will result in a reduction in opex. Again, this is counter-intuitive and undermines the validity of these models.

The signs of the first-order coefficients in the TL models provide information on the monotonicity conditions at the sample mean. However, monotonicity conditions can vary across observations in TL models. Hence, a thorough monotonicity test requires the evaluation of elasticities at each and every data point in the sample. We therefore looked in the FE (2019b) Stata files to see if we could find information on the observation–specific elasticities but were unable to find this information. Consequently, we have conducted these calculations ourselves and have summarised the results in Table 5. We observe that three of the four Ontario models have a substantial proportion of observations with monotonicity violations for CustNum: 54 per cent, 62 per cent and 100 per cent, indicating significant problems with these three models.

Table 5 **Monotonicity violations in Australian data in baseline and Ontario Translog models, percentage of observations**

	2006-2018					
Method:	**LSE**			**SFA**		
Model	**CustNum**	**CircLen**	**RMDemand**	**CustNum**	**CircLen**	**RMDemand**
Base	0%	0%	0%	0%	1%	2%
Ontario	54%	0%	0%	0%	0%	0%

	2012-2018					
Method:	**LSE**			**SFA**		
Model	**CustNum**	**CircLen**	**RMDemand**	**CustNum**	**CircLen**	**RMDemand**
Base	25%	0%	0%	0%	3%	0%
Ontario	100%	0%	0%	62%	0%	3%

Fourth, it should be noted that the Ontario data includes DNSPs that on average have a higher customer density (customers per km of network length) with a mean of 47 and a range from 10 to 83, while the Australian data has a mean of 28 and range of 4 to 76, and the New Zealand data has a mean of 12 and a range of 4 to 36. As a result, these Ontario observations play a larger role in determining the shape of the production technology in that part of the data space that is closer to the customers axis (eg the horizontal axis in Figure 2 above). If the production technology was linear, this would not be an issue. However, the production technology is non–linear and hence it is not surprising that a subset of the data which has an average customer density that differs from the remainder of the sample provides some estimated coefficients that differ to some degree. This is a consequence of the non–linear nature of the global frontier.

Fifth, as noted above, the accuracy of estimation is actually improved by having diverse characteristics in the sample. If all included DNSPs have similar characteristics then the model

will find it hard to provide robust parameter estimates. This was illustrated in Economic Insights (2014, p.28) where it was shown that the Australian data on its own exhibited insufficient variability to produce robust parameter estimates. The FE (2019a, p.36) recommendation that future benchmarking exclude international data is thus likely to remain infeasible. Similar problems are likely to accompany attempts to differentiate parameter estimates by country within the larger sample.

This leads to our sixth point that it is quite common for regulatory economic benchmarking models to include DNSPs from a diverse range of jurisdictions in the sample. Our models are unusual in allowing for differences in cost levels across jurisdictions – most do not, either because the authors have considered it unnecessary or because insufficient information has been available. And it is very rare for economic benchmarking studies to attempt to provide completely separate parameter estimates for most variables for DNSPs from different jurisdictions, again either because it was considered unnecessary or infeasible. For example, a recent economic benchmarking study prepared for the Ontario Energy Board by our fellow economic benchmarking practitioner, PEG, uses a sample of 84 DNSPs covering Ontario and all corners of the United States (PEG 2019b). That is, jurisdictions cover a wide range of operating environments from cold northern climates and large cities in Ontario, New York and Pennsylvania to warm, subtropical climates in Florida, to the sparse plains of Texas, to the mountains of Colorado, and to conurbations in temperate coastal California. Similarly, a wide range of jurisdictional regulatory regimes is included from productivity–based to incentive–based to cost of service–based. No differentiation of parameter estimates across this wide range of jurisdictional and regulatory operating environments is included.

Normality of residuals

FE (2019b, p11) state that:

> *For the validity of the SFA models, it is also a requirement that the residual term is normally distributed.*

This is incorrect. The error structure of the SFA model is the sum of a normal distribution and a time–invariant truncated normal distribution. Thus, one would expect that the residuals would actually be non–normal and hence it makes no sense to test for normality of residuals in the case of SFA. For details of the SFA methods used see:

> https://www.stata.com/manuals13/xtxtfrontier.pdf

FE (2019b, p11) also state that:

> *As an example of such a diagnostic investigation, in Figure 2 we plot the residuals of the LSE–TL model estimated over the 2006–2018 period. The residuals, which can be interpreted as percent prediction errors, are plotted against a normal distribution. If the residuals were normally distributed, the points would lie on a straight line.*

FE (2019b) do not state exactly what type of plot is being presented here. However, we will assume that it is a traditional "normal probability plot" – for a description of this method, see Levine et al (2002, p.243). The plot FE (2009b) presents is linear for the vast majority of observations (as expected) with a small number of residuals deviating from the line in the tails. While FE (2019b) has not identified these latter observations, our manipulation of the

supporting Stata and spreadsheet files provided by FE show that the 'outlier' tails are made up entirely of Ontario and New Zealand observations and hence the FE (2019b, p.11) claim of potential LSE–TL model 'misspecification' has no impact on the Australian DNSP analysis. Furthermore, we note that FE (2019b) do not present similar information for the other seven models.

In addition to this, it is unclear as to why the normality of the residuals of the LSE model is of particular interest here. The asymptotic properties of the LSE estimator do not rely upon an assumption of normality of the disturbance term. For details of the LSE method used see:

> https://www.stata.com/manuals13/xtxtpcse.pdf

Finally, on closer inspection of the FE Stata file "Fig2 QQ plot.do" we find that the following code has been used:

```
xtpcse lvc ly2-ly4 ly22 ly23 ly24 ly33 ly34 ly44 lz1 yr cd2 cd3 d2-d13, c(a) het
* Q-Q plot
predict yhat_TL
gen e_TL = lvc - yhat_TL
gen res_LSE_TL = 100*(exp(e_TL) -1)
qnorm res_LSE_TL
```

This Stata code indicates that a linear transformation of the exponent of the residuals has been plotted and not the actual residuals themselves. If the residuals of this regression equation where normally distributed, then an exponential transformation would produce a series that has a log-normal distribution by definition. Hence, in our assessment this QQ plot analysis of the exponent of the residuals is clearly invalid.

Tests of TL versus CD models

FE (2019b, p12) state that:

> *It is also possible to undertake statistical tests to evaluate the comparative fit of the models to the data. One such test is a test of the fit of the TL model versus the CD model, which is a special case of the TL model. We have carried out this test, and the results show that, in all four cases—LSE long sample, SFA long sample, LSE short sample and SFA short sample—the test rejects the hypothesis that the CD model is an acceptable simplification of the TL model. (footnote 25)*
>
> *We do not wish to imply that there is no value in estimating the CD model. However, the above arguments strongly suggest that, for sound statistical reasons, the results of the econometric estimations need to be treated with appropriate caution.*

The FE (2019b) test results are reported in their footnote 25. Again, it is not clear what type of hypothesis test has been conducted. It is not indicated whether it is a Likelihood Ratio test or a Wald test or another test. However, given that the testing methods are standard, the reported results indicate that the CD should be rejected in favour of the TL on this basis. This result is not surprising and would agree with similar hypothesis test results presented in Economic Insights reports in the past. For example, Economic Insights (2014, p36) found that while all of the six second-order output coefficients in the TL model were individually statistically insignificant, collectively they were significant. This can be explained by the fact

that the squares and cross products in the TL model are highly correlated leading to large standard errors on these estimated coefficients and hence low (individual) t-ratios.

In most cases, when one has a sufficiently large data set, one would expect a statistical test to indicate that the TL is a better fit to the data relative to the CD, since the TL is a second-order approximation to an arbitrary functional form while the CD is a more restrictive first-order approximation. However, the added flexibility of the TL also lends it towards a greater propensity to obtain monotonicity violations for some data points, which are problematic as discussed above. There are trade-offs in every modelling decision that is made. Hence, we have chosen to report both TL and CD models in our benchmarking reports for the AER.

Monotonicity violations

As discussed in Economic Insights (2014, pp. 32–33), it is important that econometric opex cost function models satisfy the technical requirement that an increase in output can only be achieved with an increase in cost – this is known as the monotonicity requirement. It is an important economic requirement. In simple terms, it is the requirement that there are no free lunches. If it is not satisfied, it implies that DNSPs could produce more output without any additional cost or, if the cost elasticity is negative, at less cost – something that does not reflect engineering reality. Because the translog models include second order terms, it is necessary to check that the estimated cost elasticities for each output are positive at each observation.

FE (2019b, p16) state that:

> *We have previously put forward arguments why minor monotonicity violations should not disqualify a translog model from being used to assess the efficient base year level of opex for DNSPs. These violations need to be considered in the context of the uses that the AER makes of the econometric models, the main uses being:*
>
> > *• to obtain estimates of the DNSPs' relative efficiencies; and*
> >
> > *• to obtain output weights to enable the calculation of the combined output growth across the three main output variables for use in its roll–forward methodology.*
>
> *Neither of these applications of the econometric models involve calculating elasticities for specific observations.*

We do not agree with this statement. The panel data SFA model implemented in Stata is that proposed by Battese and Coelli (1988). The formula used for the calculation of efficiency scores is also presented in Battese and Coelli (1988) where it is clearly shown that the efficiency scores calculations incorporate the residuals of the estimated model, where these residuals are calculated at each data point and thus explicitly make use of the localised slope information and hence implicitly make use of the localised elasticities as well. As a result, if an output elasticity is negative for one particular output (eg lines), this would imply that a small increase in the production of this output, with everything else held constant, would result in a reduction in the calculated value of the residual and hence a reduction in the calculated efficiency score. That is, an increase in output produces a reduction in efficiency - this is further illustrated below in Figure 3. This is a very unusual property for an efficiency measure and not one that we would recommend.

The property of monotonicity appears not to be well understood by many people. It is perhaps useful to spend some time explaining the importance of the monotonicity property in any efficiency analysis that involves frontier estimation.

First, it must be emphasised that a negative production elasticity implies a production possibility curve with an incorrect (positive) slope. To illustrate this, consider a simple cost frontier with two outputs: customers and line length. The localised part of a Translog cost frontier can be approximated by the following Cobb Douglas function in log form:

$$\ln(opex) = \beta_0 + \beta_1 \ln(lines) + \beta_2 \ln(cust)$$

Rearranging the equation to have ln(*lines*) on the LHS we obtain:

$$\ln(lines) = \ln(opex) - [\beta_0 + \beta_2 \ln(cust)]/\beta_1$$

Taking derivatives with respect to ln(*cust*) we obtain:

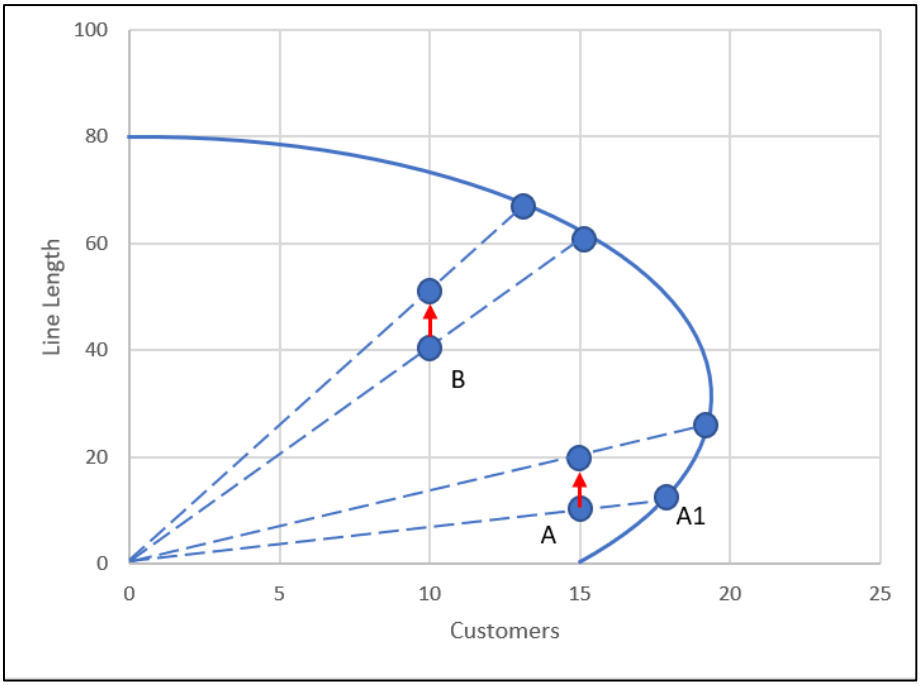
$$\frac{\partial \ln(lines)}{\partial \ln(cust)} = -\frac{\beta_2}{\beta_1}$$

Using the log differentiation rule this is equivalent to:

$$\frac{\partial lines}{\partial cust} = -\frac{\beta_2}{\beta_1}\frac{cust}{lines}$$

This derivative is the slope of the production possibility curve. It will be negative (as expected) if the two output elasticities are positive. However, this slope will be positive if one of these two elasticities is negative. An example of this is provided in Figure 3 below, where a production possibility curve (for a particular level of opex) is drawn to have a negative slope for one section of the curve and then a positive slope for the remaining section.

Figure 3 **Effect of monotonicity violations on efficiency scores**



Data for two DNSPs (A and B) are drawn on the diagram. DNSP A is projecting onto part of the frontier which has an incorrect positive slope while DNSP B is projecting onto part of the frontier which has a correct negative slope. Cost efficiency is measured along the ray from the origin to the frontier surface. For example, DNSP A has a cost efficiency score equal to the ratio 0A/0A1, or approximately 0.85. That is, it is producing 85 per cent of its potential output.

Next, we increase the line length output by 10 units for both DNSPs, while holding customers and opex constant. Normally, one would expect that this should have the effect of increasing the efficiency scores of both DNSPs. This is true for the case of DNSP B, but conversely, we see that the efficiency score of DNSP A decreases when the output is increased. This is because the production possibility curve has an incorrect positive slope for this DNSP.

We hence do not recommend that efficiency scores be calculated for a DNSP where it has a production elasticity that is negative.

Bootstrapping

FE (2019, p55) present the following quote from Prof Coelli in ACCC (2012, p113) as saying:

> *I would suggest that the construction of bootstrap confidence intervals for DEA efficiency scores could provide some useful information regarding the degree to which these DEA results obtained from small samples can be relied upon.*

This quote is technically accurate. However, the interpretation of this sentence becomes clearer when the full quote listed in ACCC (2012, p113) is actually provided:

> *Given that DEA frontiers are arguably more flexible than a second–order parametric frontier, such as the Translog, one would expect that the data requirements for DEA are greater than those of SFA. Hence, I believe that the existing rules of thumb used in the DEA literature are generally too low. I would suggest that the construction of bootstrap confidence intervals for DEA efficiency scores could provide some useful information regarding the degree to which these DEA results obtained from small samples can be relied upon.*

In this fuller context it is apparent that the quote was not actually a ringing endorsement of using bootstrapping in efficiency models. Instead the quote relates to a discussion of how some of the commonly used rules of thumb for determining how much data is needed for a DEA model are questionable, and that given that DEA is a non–statistical (deterministic) method, bootstrapping might provide some insight into how unreliable the DEA method is when small samples are used.

A more complete view of Prof Coelli's views on bootstrapping can be found in Coelli et al (2005, pp202–203) where, in a discussion of DEA methods, it is stated that the authors have "a number of reservations regarding the bootstrap". Furthermore they state that "it does not make much sense for one to apply bootstrapping methods to a DEA analysis based upon census data", where census data refers to the situation where the data under study is not a random sample from a population and instead is a full census of the population under study. This is clearly the case in this AER analysis at hand, since the data on DNSPs is a census and not a sample. Bootstrapping is a resampling technique that is useful when attempting to assess the

influence of sampling variability on an estimated model. If the data is a census then clearly no sampling is involved in selecting the data set.

It should also be noted that in no other part of the Coelli et al (2005) book is bootstrapping discussed or recommended for use with SFA or any other frontier technique.

The main application of bootstrapping is to help construct confidence intervals around point estimates when the underlying distribution is unknown. Therefore, it has been commonly used in deterministic approaches (such as DEA) which do not account for statistical uncertainties. As a non-parametric alternative to the parametric approach to constructing confidence intervals, the residual resampling in a regression assumes that the residuals are independent and identically distributed, that is, the distribution of errors around the regression line is the same for all values of the independent variable. This assumption may not be true or may be inconsistent with the assumed error term in conventional regression analysis and/or frontier modelling which typically allows for autocorrelation, heteroskedasticity and within group correlation. As a result, it is not common in regression analysis to construct bootstrap-based confidence intervals.

It should be emphasised that Economic Insights (2019) and our earlier economic benchmarking reports for the AER have all provided information on the statistical reliability of the efficiency scores derived from their SFA and LSE models. These are provided in the form of asymptotic standard errors associated with the DNSP–level dummy variables in the LSE models and confidence intervals on the efficiency predictions from the SFA models. This information is clearly presented in the Stata output files attached to the main report each year. The AER has regularly taken note of this type of information in making its deliberations regarding the reliability of the information derived from these estimated models.

We also observe that the application of the bootstrap in parametric frontier models (eg SFA or LSE) is not commonly done in practice. A few studies have attempted to investigate the properties of the bootstrap in these models, with limited success. For example, see Kim et al (2007) who conduct a study entitled: "The accuracy of bootstrap confidence intervals for efficiency levels in stochastic frontier models with panel data". They look at a number of alternative methods for constructing confidence intervals and in their conclusions state that (Kim et al 2007, p180): "It should be remembered that all of these methods are valid only for large T, and when T is not large enough that the identity of the best firm is clear, none of them will really be reliable".

The Kim et al (2007) study also clearly illustrates that the implementation of the bootstrap in frontier models involves a number of complexities and potential biases and cannot be implemented in a simple manner. The data generating process associated with frontier models is complex and hence requires careful consideration when implementing bootstrap methods.

It is not clear to us that the FE (2019b) implementation of the bootstrap technique takes into account these complexities. The bootstrap procedure they describe appears to be a very simple process of sampling with replacement from an empirical distribution. This appears to be insufficient given the complexities of the models at hand.

To begin with we consider the SFA model which involves an error distribution that is the sum of two random variables – one with a normal distribution (v_{it}) and another with a time–

invariant truncated normal distribution (u_i). It is not clear that the FE bootstrap procedure has accommodated this composed error structure in an appropriate manner. First, it appears that the FE bootstrap procedure has implicitly assumed that the random variables u_i are fixed parameters when generating their 1,000 samples, but then contrary to this they assume the u_i are random variables when estimating the SFA model in each replication, creating a clear inconsistency between the data generating process and the model structure itself. Second, in the FE bootstrap data generating process the u_i are predicted using the Stata conditional expectation predictor $E(u_i|\varepsilon_i)$, where ε_i is the T×1 vector of the $\varepsilon_{it} = v_{it} + u_i$, while the efficiency scores are then predicted using $TE_i = E[\exp(u_i|\varepsilon_i)]$, which is not equivalent to $\exp[E(u_i|\varepsilon_i)]$, as is explained in Battese and Coelli (1988).[5] Once again, this creates a clear inconsistency between the data generating process and the model structure itself. Third, it is not clear how this FE bootstrap procedure has accommodated the widely known downward bias problem that is inherent in bootstrap analyses of frontier models, which is discussed in Kim et al (2007) and elsewhere.

Next, consider the LSE model, which involves an error term characterised by autocorrelation, heteroskedasticity and within group (ie within DNSP) correlation. First, it is apparent that the FE bootstrap procedure has not accommodated these three aspects of the assumed data generating process. Second, it is again not clear how this FE bootstrap procedure has accommodated the widely known downward bias problem that is inherent in bootstrap analyses of frontier models, which is discussed in Kim et al (2007) and elsewhere.

In footnote 101 on page 56 FE (2019b) state:

> *The bootstrapping approach assumes that the residuals can be treated as independent, identically distributed random variables. For a well–specified regression model without autocorrelation and heteroscedasticity, this assumption holds in large samples. Intuitively, the rationale behind bootstrapping is that, if the residual terms are random draws for the same distribution, then any residual could, with equal probability, be associated with any of the observations.*

This statement tends to confirm our suspicion that this bootstrap analysis is most likely flawed.

In addition to these points above, we also have particular reservations regarding the practical application of the bootstrap technique to TL functional forms, in particular in regards to monotonicity violations. The bootstrap procedure involves the estimation of 1,000 model replications for each of the four TL models considered. From our inspection of the FE Stata code we can find no information on the calculation of monotonicity tests at each observation of each model in each of these 1,000 replications. We would suspect that monotonicity violations would be common in the TL models.

Our suspicions in this regard appear to be confirmed by the results presented in Figure 11 of FE (2019b) where it is clearly shown that the point estimates for the short sample SFATL models for both Ergon and Energex do not lie within the 90% confidence intervals obtained. This is most unusual and should have immediately rung serious alarm bells for the analysts responsible for running this bootstrap analysis. This result suggests that many of the underlying

[5] Battese and Coelli (1988) show this for the production frontier case which is easily transferred to this cost frontier case.

sets of 1,000 model replications are most likely very unstable and unreliable for these TL models.

Finally, FE (2019b) appears to argue that where the proposed opex sits within the confidence interval, then there is no evidence of material inefficiency. However, in regulatory applications, the confidence interval has not been used to set range of possible efficient values. Rather, it is a statistical construct used to estimate precision of the point estimate (eg the width of the confidence interval and the precision of the point estimate will generally be negatively related to the sample size). The point estimate provides the best estimate about the unknown true efficient value, while none of the other values within the confidence interval do. Confidence intervals may be useful in informing the degree of confidence in the point estimate, and thus the weights to apply to the estimate when multiple estimates from different sources/methods are available. They do not mean that all values within the confidence interval can be viewed as being efficient.

We also note that in constructing the confidence interval for base year target opex, FE (2019b) appears to apply the 0.75 efficiency target. We note that the AER uses this lower efficiency target (rather than either 1.00 or the score of the most efficient DNSP) to provide an error margin for data/modelling uncertainties and thus the FE (2019b) approach involves double counting.

OEF adjustment methods

FE (2019b, p49) criticises the AER draft decision for not adopting the following recommendations from FE (2019a) regarding OEF inclusion:

- *investigating the inclusion of additional cost driver variables in its model, which should become more feasible over time as the benchmarking sample size increases; and*
- *making ex-ante adjustments for any costs associated with OEFs that are unexplained, or poorly explained, by the cost driver variables that are included in the model—as Ofgem does.*

With regard to the first point, the ability to include additional OEF variables in the models directly is limited by the availability of relevant data for both the Australian and overseas DNSPs included in the database. While the AER has scope to require the provision of relevant data for Australian DNSPs, it is not able to force the overseas DNSPs included in the database to provide this information. Consequently, if these variables are not available for the overseas DNSPs, the direct inclusion of the variables in the models is not possible. As highlighted above, the lack of data variability across the Australian DNSPs means that extension of the time series length is unlikely to significantly improve the ability to obtain robust parameter estimates based on Australian data only. And, it should also be noted that degrees of freedom considerations and correlation among exogenous variables in regressions limit the number of operating environment variables that can usefully be included directly in economic benchmarking models. The prospects of being able to directly include additional OEF variables directly in the

models are thus quite limited. This makes the use of subsequent adjustment the only way of allowing a fuller treatment of operating environment factors.

Moving to the second point, there are four ways to allow for OEF differences between DNSPs in economic benchmarking models, all of which Economic Insights and the AER use in the DNSP economic benchmarking analysis:

1. ex–ante changes to data – ex–ante changes are made to ensure similar coverage of Australian DNSP activities by, for example, removing costs associated with metering, connections, public lighting, fee–based and quoted services and opex associated with solar feed in tariffs
2. direct incorporation of OEFs in the model – key network densities and proportion of undergrounding are included in the models
3. ex–post adjustment of results for additional OEFs – ex–post adjustments are made to account for additional OEFs not directly captured in the models, and
4. second stage regression analysis – Economic Insights (2014) used second stage regressions as a check on the need for an additional allowance for OEFs in the MPFP model.

All four OEF adjustment methods have their own advantages and disadvantages, and the choice of which one (or ones) is best to use depends on the individual circumstances. As noted with regard to the first point, the limited ability to obtain additional data on OEFs for the overseas DNSPs limits the scope to make ex–ante data changes to allow for additional OEFs, just as it limits the scope to include additional OEFs directly in the models. But, where data for particular OEF variables are not universally and consistently available across countries but are available for Australian DNSPs, ex–post adjustment for those factors can be made when comparing Australian DNSP performance to the most efficient Australian firms.

It should also be noted that in the example referred to in FE (2019b), Ofgem works with a sample of only 14 domestic DNSPs and, correspondingly, estimates models that include a minimal number of exogenous variables – in most cases only one variable although this is sometimes a constructed ‘composite’ variable. It excludes costs for some DNSPs associated with unusual operating environments, undertakes its modelling and then adds back in its view of efficient costs for the excluded items. This process potentially introduces scope for DNSP gaming regarding the size of costs to be excluded and arbitrariness in regard to the regulator’s view of efficient costs for those items. And, limiting the sample to national data only requires the estimation of models with minimal detail, likely because attempts to estimate more detailed models would run into the similar lack of data variability issues across DNSPs as we have found in Australia.

We also note that, in addition to undertaking ex–ante data adjustments, Ofgem does make additional ex–post adjustments by using DNSP–specific ‘special factors’ for some DNSPs and the Norwegian regulator (NVE) makes post modelling adjustments using second stage regressions before arriving at its decision. We note that FE (2015, p.77) earlier arrived at the following conclusion:

> *AER may need to adopt a hybrid of the Ofgem approach (which involves more direct and bespoke scrutiny of the companies), but taking account of a broader range of factors, as does NVE.*

This is, in fact, consistent with the approach adopted in Economic Insights (2014) and subsequent AER decisions.

The challenge with economic benchmarking for regulatory purposes is to determine how much of the unexplained residual from modelling to allocate to DSNP inefficiency and how much to latent (or unobservable) heterogeneity among the included DNSPs. Assuming all is inefficiency likely provides an upper bound for base year cost adjustments while assuming it is due to latent heterogeneity will provide a lower bound for base year adjustments. The former may produce too large an adjustment while the latter will almost certainly produce too low an adjustment. Our use of the two–step process for calculating the overall adjustment for operating environment differences and a combination of ex–ante data adjustment and ex–post additional OEF adjustment along with models that include more exogenous variables provides a means of reaching the most appropriate point within this range of possible base year adjustments. It also allows the impact of more operating environment factors to be adjusted for than have earlier economic benchmarking studies.

Finally, FE (2019b, p50) includes a quote from the Australian Competition Tribunal which appears to criticise the use of ex–post OEF adjustment as being inadequate to overcome the effect of including 'non–comparable' data in the economic benchmarking models. While it is not clear what the Tribunal had in mind by 'non–comparable' data, we assume it was referring to the inclusion of overseas as well as Australian DNSP data and an apparent acceptance of arguments that the data across countries do not meet 'poolability' tests. However, as noted above, such 'poolability' tests are based on flawed logic when there is inadequate data variability among the Australian DNSPs to produce robust parameter estimates and there are firm grounds to believe that the basic cost relationships across DNSPs are similar, despite possible differences in cost levels due to things such as more extreme winter weather conditions. There is thus no evidence that the data are 'non–comparable'.

References

Australian Competition and Consumer Commission/Australian Energy Regulator (ACCC/AER) (2012), *Benchmarking Opex and Capex in Energy Networks*, Working Paper No. 6, Melbourne, May.

Australian Energy Regulator (AER) (2015), *Final Decision: Ausgrid Distribution Determination 2014–15 to 2018–19 – Attachment 7: Operating Expenditure*, Melbourne, April.

Battese, G.E., and T.J. Coelli (1988), "Prediction of Firm–Level Technical Efficiencies with a Generalised Frontier Production Function and Panel Data", *Journal of Econometrics*, 38, 387–399.

Coelli, T, D.S.P. Rao, C. O'Donnell and G. Battese (2005), *An Introduction to Efficiency and Productivity Analysis*, 2nd Edition, Springer.

Economic Insights (2014), *Economic Benchmarking Assessment of Operating Expenditure for NSW and ACT Electricity DNSPs*, Report prepared by Denis Lawrence, Tim Coelli and John Kain for the Australian Energy Regulator, Eden, 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, Eden, 22 April.

Economic Insights (2019), *Economic Benchmarking Results for the Australian Energy Regulator's 2019 DNSP Benchmarking Report*, Report prepared by Denis Lawrence, Tim Coelli and John Kain for the Australian Energy Regulator, Eden, 5 September.

Frontier Economics (FE) (2015), *Taking Account of Heterogeneity between Networks When Conducting Economic Benchmarking Analysis*, Report prepared for Ergon Energy, February.

Frontier Economics (FE) (2019a), *AER Benchmarking*, Report Prepared for Energy Queensland, 15 January.

Frontier Economics (FE) (2019b), *Assessment of the AER's Benchmarking Analysis*, Report Prepared for Ergon Energy And Energex, 9 December.

Kim, M., Kim, Y. and Schmidt, P. (2007), "On the accuracy of bootstrap confidence intervals for efficiency levels in stochastic frontier models with panel data", *Journal of Productivity Analysis*, 28, 165–181.

Levine, D., Stephan, D., Krehbiel, T. and Berenson, M. (2002), *Statistics for Managers Using Microsoft Excel*, Prentice Hall, New Jersey.

Pacific Economics Group (PEG) (2008), *Benchmarking the Costs of Ontario Power Distributors*, Report prepared for Ontario Energy Board, Madison, 20 March.

Pacific Economics Group Research, LLC (PEG) (2019a), *Empirical Research in Support of Incentive Rate–Setting: 2018 Benchmarking Update*, Report prepared for Ontario Energy Board, Madison, August.

Pacific Economics Group Research, LLC (PEG) (2019b), *IRM Design for Toronto Hydro–Electric System*, Report prepared for Ontario Energy Board, Madison, 22 May.