



Σ ECONOMIC
i INSIGHTS Pty
Ltd

Domestic Transmission Capacity Services Benchmarking Model

Final Report prepared for
Australian Competition and Consumer Commission

John Fallon, Michael Cunningham, and Tim Coelli

1 September 2015

Economic Insights Pty Ltd

10 By Street, Eden, NSW 2551, Australia, AUSTRALIA

Ph 0419 171 634

WEB www.economicinsights.com.au

ABN 52 060 723 631

CONTENTS

1	Introduction.....	1
1.1	Terms of reference	1
1.2	Context.....	2
1.2.1	ACCC’s 2012 determination.....	3
1.2.2	ACCC’s 2014 inquiry	4
1.3	Consultation	4
1.4	Outline of the Report	5
2	Description of 2011 and 2014 Data	6
2.1	Dimensions	6
2.2	Providers	6
2.3	Routes & ESAs	6
2.4	Key Contract Parameters	8
2.5	Quality of Service Indicators	13
2.5.1	Protection	13
2.5.2	Interface	13
2.5.3	Quality of service scores	14
2.6	Indicators of market size and competition	16
3	Exploratory Data Analysis.....	19
3.1	Capacity and Distance.....	19
3.2	Contract start date and term	20
3.3	Conditioning variables.....	21
3.4	Categorical Variables.....	26
4	Review of 2012 DTCS Econometric Model.....	28
4.1	The 2012 econometric model.....	28
4.2	Diagnostic tests	28
4.3	Application to 2014 data.....	31
4.4	Estimating with the pooled 2011 and 2014 data.....	33
5	Developing a Preferred Model.....	34
5.1	Methodology	34
5.1.1	Economic & Econometric Specification.....	35
5.1.2	Variables and Expected Signs.....	37

5.1.3	Estimation method	39
5.2	Deriving the preferred model.....	43
5.2.1	Submissions to Draft Report.....	43
5.2.2	Addressing issues raised in the submissions.....	45
5.3	Preferred models	46
5.4	Concluding comments	51
6	Applying the models to regulated Routes.....	52
6.1	Current pricing model.....	52
6.2	Pricing models	52
6.2.1	Application of Model 2.....	53
6.2.2	Application of Model 3.....	54
6.2.3	Rate of change in prices.....	55
6.3	Tail-end pricing.....	55
6.4	Comparison of pricing benchmarks	58
6.5	Allowances for productivity	61
6.6	Setting prices based on the mean or some other percentile	61
	References.....	62
	Annex A: Additional Demand and Supply variables.....	63
	Annex B: Data Management Documentation.....	66
	Annex C: Review of the 2012 Model	67
	Annex D: Econometric Specification Search	78

1 INTRODUCTION

The Australian Competition and Consumer Commission (ACCC) has contracted Economic Insights Pty Ltd ('Economic Insights') to provide advice and undertake econometric modelling in relation to prices of Domestic Transmission Capacity Services (DTCS).

The objective of this study is to produce a model for determining DTCS prices on regulated routes, based on a regression model that provides the best explanation of observed commercial prices on routes that have been classified as competitive by the ACCC. With sufficient recognition of the factors that give rise to different prices for different contract specifications, this method should lead to reasonable benchmarks for setting prices for regulated DTCS services.

1.1 Terms of reference

The ACCC intends to set regulated DTCS prices in the 2015 DTCS FAD using a benchmarking approach. The benchmark prices are to be derived from a regression model based on price data from unregulated routes, with appropriate adjustments.

The deliverables of this project include, firstly, a presentation to ACCC staff which sets out a preliminary analysis of the:

- (a) differences between the data sets used for the 2012 and 2015 FAD
- (b) applicability of the 2012 regression model to the 2015 FAD
- (c) appropriateness of including new variables in the model according to economic principles, including the cost and market drivers of transmission service prices.

The second set of deliverables is the draft and final versions of report provided to the ACCC. The report will be informed by consultation with stakeholders and with technical experts engaged by them via a one-day workshop, and the report will include consideration of the comments of stakeholders and stakeholder technical experts. The report must provide:

- (a) Details of the recommended model to calculate DTCS prices on declared routes, based on regression equations which best fit observed prices on competitive routes.
- (b) Explanation of any adjustments required to set efficient prices in declared routes to allow for differences in the characteristics of declared areas from competitive routes.
- (c) Justification of the form of equations chosen, with explanation of alternative forms proposed or considered and any qualitative considerations supporting the approach.
- (d) Justification for the explanatory variables chosen, with explanation of other variables considered or proposed, including consideration of the following variables: data rate, distance, region or route type, protection, technology type, quality of service, measures of demand and competition.
- (e) Statistical analysis and explanation of the input data and model outputs including:

-
- (i) summary statistics for the input variable data and model outputs / variables in the estimation data (such as mean and median, standard deviation, maximum and minimum, scatter plots), coefficient standard errors and post-estimation diagnostic tests
 - (ii) differences in price outputs between the 2012 FAD model and the recommended model for the 2015 FAD
 - (iii) differences between the 2012 FAD and 2015 FAD data sets which affect the form of the recommended model for the 2015 FAD
 - (f) Explanation of any allowance in the model for change in prices over the 2015 FAD period to reflect expected productivity and cost movements.
 - (g) Explanation of the operation and use of the model, with advice on interpreting its outputs.
 - (h) The methodology used to identify and treat outliers in the data set.
 - (i) Recommendations on whether the regulated prices should be based on the mean of the estimated data or some other percentile.

1.2 Context

As part of its responsibilities for administering the telecommunications access regime, the ACCC is required to determine prices for access to declared telecommunication wholesale services and infrastructure. The DTCS is the declared wholesale transmission service under s. 152AL of the *Competition and Consumer Act 2010* (CCA). It was first declared in 1997 and in March 2014 the ACCC extended the declaration until 2019. The declaration is limited to transmission services that meet the DTCS service description on routes and in areas that the ACCC determines are not competitive.

The term ‘transmission’ refers to high capacity data links that are used to carry large volumes of communications traffic such as voice, data or video communications. DTCS services include transmission services that are high capacity (2Mbps and above), permanent, symmetric and uncontended. The DTCS has been described as:

a high capacity transmission service differentiated from other transmission services on the basis that it:

- *is a wholesale input into the provision of other services and not a resale service. That is, the DTCS service must be used in combination with an access seeker’s infrastructure to provide other end-to-end services*
- *is a point-to-point service*
- *may be provided over a number of transmission mediums including copper, fibre and microwave*
- *is a high capacity service acquired at different data rates above 2 megabits per second (Mbps)*
- *is symmetric, that is it has the same data rate in both directions, and*

- *is an uncontended service - this means that the capacity of the service is dedicated to one access seeker only and not shared.* (ACCC 2014c p.18)

The ACCC has maintained DTCS regulation on service routes where competition is assessed to be ineffective or where access to the DTCS is limited. To date, the following services have been deregulated: inter-capital transmission between Brisbane, Sydney, Canberra, Melbourne, Adelaide and Perth; 23 regional routes; and 88 metropolitan inter-exchange routes. From early 2015 an additional 112 metropolitan inter-exchange routes and an additional (net) 5 regional routes became undeclared services.¹

The ACCC's first final access determination (FAD) for the DTCS was in June 2012. The price and non-price terms of access determined by the ACCC are intended to provide guidance to parties when negotiating access agreements for a DTCS. Parties are free to agree other terms and conditions of access, but if there is no explicit agreement on other terms and conditions, those determined by the ACCC apply. To determine prices for the 2012 FAD, the ACCC used its 'domestic benchmarking approach', in which benchmark prices from the competitive (deregulated) service routes are used to derive prices that would be likely to apply in the regulated service routes if they were competitive.

1.2.1 ACCC's 2012 determination

The regulated DTCS prices adopted in 2012 for declared areas and routes were based on modelling the deregulated prices charged in competitive service routes against relevant service attributes. The competitive prices provided a guide to the appropriate regulated prices for similar services.

Data including prices for DTCS products in competitive areas and routes were collected by the ACCC from 7 providers in 2011. The dataset comprised approximately 4,500 records representing individual contracts (Data Analysis Australia (DAA) 2012). The price is defined as the charge (in \$) for an individual service supplied for a period of 12 months. Connection charges are an additional one-off charge for connecting to the service.

Corresponding to each price are the attributes of the service to which it relates. The data collected by the ACCC included: 'data rate' measured in millions of bits per second (Mbps); distance of the service route in kilometres (km); route category (see footnote 1); distance category; provider (the network owner); path protection or redundancy status (whether another route is available in case of failure or constraint); network interface type (SDH/Ethernet); and contract term.

Linear regression was used to relate the log values of prices of the DTCS products to log values of service attributes, particularly carriage distance, data rate, route category and protection status (i.e. whether there is redundancy in case of failure).² The ACCC also

¹ The route categories are: inter-capital transmission (between two capital cities); regional routes (capital-regional or regional-regional); metropolitan inter-exchange (within a single capital city between Exchange Serving Areas); and regional and metropolitan tail-end transmission within a single Exchange Serving Area (ESA)).

² No services in the tail-end route category were included in the dataset because all of these remained declared.

developed a service quality indicator variable, which it used in the model, in recognition of the fact that providers differ in the service quality and reliability their networks provide.

Benchmark DTCS prices could then be determined using the predictions from the model. The ACCC used a 'mean value approach', meaning that it based prices on the model's predictions (the conditional expectation of the competitive price) and not using an upper percentile. However, it made an upward adjustment by adopting the benchmark price of a provider with high service quality. No price escalation formula was included in the determination.³

Connection charges were differentiated between data rate and network interface (Ethernet and SDH) and based on the averages of connection charges in the data sample within each data rate and interface classification.

1.2.2 ACCC's 2014 inquiry

The current inquiry into making a new DTCS FAD was initiated in May 2014,⁴ and includes two consultation processes, one for pricing and the other for non-price terms and conditions. This report contributes to the pricing aspect of the inquiry.

The ACCC released a position paper on pricing methodology in November 2014 and has indicated that it intends to retain the domestic benchmarking approach in its current inquiry, but will carefully consider opportunities to improve the regression model to ensure it is effective, and how the model can best be used to determine efficient prices (ACCC 2014a; b). The econometric analysis will have regard to data previously collected in 2011 and further data collected in 2014, including access agreements entered into in the intervening years.

The ACCC indicated it will engage stakeholders during the modelling process, including in relation to the preliminary analysis of the dataset to be used in the regression model (subject to suitable confidentiality arrangements) and will support the use by stakeholders of independent experts to carry out their own analysis using the same dataset the ACCC will use. It will also engage stakeholders in regard to the most appropriate way to determine final prices from the outputs of the regression model (ACCC 2014a). Some of the issues to be considered in regard to the application of the model for pricing regulated services include: whether to use the 'mean value approach' used in 2012; and whether prices should decrease over the access period to reflect technical change.

1.3 Consultation

This final report has benefitted from comment by industry participants and technical experts engaged by them. It follows an earlier workshop paper initially prepared by Economic Insights (2015b) for workshops with industry stakeholders and technical experts held on 24

Hobart and Darwin were classified as regional rather than capital cities.

³ See DTCS pricing calculator. Prices based on the econometric model were upwardly adjusted by 10.2% to reflect the price of a high quality service provider. Prices for tail-end route services assume the distance is 2 km. A 40% uplift was applied to routes via the Bass Strait.

⁴ The 2012 determination would have expired on 31 December 2014, but on 5 November 2014 the ACCC extended it until a new determination comes into force.

April 2015. Initial comments were provided by stakeholders at the workshop and in written submissions made at around that time. A draft report prepared by Economic Insights (2015a) was made available to the same stakeholders and experts and several written submissions were received in response. The observations and arguments presented in these submissions have been carefully considered in preparing this report.

1.4 Outline of the Report

Chapter 2 of this paper provides a brief summary of the datasets collected by the ACCC from DTCS service providers in 2011 and 2014, which comprise records of individual contracts for wholesale transmission services on regulated and unregulated routes across Australia.

Chapter 3 highlights some key points from the exploratory data analysis.

Chapter 4 discusses the review of the suitability for the ACCC's present transmission pricing process of the econometric model used in the 2012 DTCS FAD. This chapter replicates the 2012 model using the 2011 dataset, tests some alternative specifications, and then applies these models to the 2014 data set and assesses their performance.

Section 5 summarises the methodologies and procedure for developing the preferred models. Some of the main issues and perspectives put forward by the industry stakeholders and technical experts are summarised. This chapter also presents preferred econometric models.

Section 6 discusses how the preferred model for price formation on unregulated routes can be applied to determine benchmark prices for regulated rates.

Annex A contains a list of additional demand and supply variables provided by the ACCC.

Annex B provides data management documentation.

Annex C is a detailed review of the applicability of the 2012 DTCS FAD econometric model to the current review.

Annex D sets out detail of the research process for deriving the preferred model presented in this report.

2 DESCRIPTION OF 2011 AND 2014 DATA

Datasets were provided by the ACCC for 2011 and 2014, each consisting of a snapshot of all individual contracts supplied by each DTCS transmission service provider to each of its customers on regulated and deregulated routes. These snapshots were obtained in January 2011 and November 2014. The 2011 dataset was used for the 2012 DTCS FAD.

This section compares the 2011 and 2014 data. Most of the econometric analysis in this report uses the 2014 dataset, although some initial analysis is carried out with the 2011 data and with the two datasets combined.

2.1 Dimensions

The 2011 dataset contains 13,470 records in total, comprising 9,375 services (or approximately 70 per cent) on regulated routes and 4,095 (or about 30 per cent) on deregulated routes. The 2014 dataset has 18,247 records, including 11,480 services on regulated routes (or approximately 63 per cent) and 6,767 services (or about 37 per cent) on deregulated routes. The records relating to deregulated routes are the most important for this study, since they form the data sample used in the econometric analysis.

2.2 Providers

Seven different service providers are represented in the 2011 data and nine different providers in the 2014 data.⁵ The four largest DTCS providers (by number of services) together comprised 96.2 per cent of contracts on regulated routes, 85.8 per cent of contracts on deregulated routes and 92.4 per cent of all contracts in 2014.

_____. Smaller players have increased their market shares over the two periods.

2.3 Routes & ESAs

The number of routes has increased considerably between the two periods, from 3,295 to 4,986 distinct routes, mainly reflecting the inclusion of tail-end routes, all of which are regulated and so are not included in the econometric analysis. In the 2014 dataset there are 1,589 deregulated routes. There are more B-end ESAs than there are A-end ESAs, reflecting the general hub-spoke pattern of the network, based on the capital cities in each state. The 'hub and spoke' network design is indicated by the fact that, almost all regional routes have A and B ends within the same state, and in almost all cases where A and B ends are in different states the route is classified as inter-capital.

■ _____

In relation to route categories (inter-capital, metropolitan, regional, and tail-end), Table 2.1 shows data for the numbers of services on each route type in each of 2011-12 and 2014-15. Some points to note in relation to this data are as follows.

- There are far fewer services on regulated metropolitan routes and many more services on deregulated metropolitan routes in 2014 than in 2011.
- 72 per cent of deregulated services in 2014 were for metropolitan routes of 1 km or more, 6 per cent were for metropolitan routes less than 1 km, 14 per cent were for inter-capital routes, and 9 per cent were for regional routes.
- In 2011, 95 per cent of the deregulated routes had distances of less than 732 km, and the average distance was 144 km. This was longer than on the regulated routes, where the average distance was 67km. In 2014 the average distance on deregulated routes was 176 km and the average distance on regulated routes was 107 km. The 2014 review of the declaration of DTCS services resulted in the deregulation of a number of routes (ACCC 2014c).

Table 2.1: **Route classes 2011 & 2014, frequencies**

	Regulated		Deregulated		Total	
	No.	%	No.	%	No.	%
2011						
Inter-capital			524	12.8		
Metro ≥ 1 km			2,823	68.9		
Metro < 1 km			320	7.8		
<i>Metro sub-total</i>			3,143	76.8		
Regional			428	10.5		
Total	9,375	100.0	4,095	100.0	13,470	100.0
2014						
Inter-capital			942	13.9		
Metro ≥ 1 km			4,857	71.8		
Metro < 1 km			387	5.7		
<i>Sub-total Metro</i>			5,244	77.5		
Regional			581	8.6		
Tail-end			-	-		
Total	11,480	100.0	6,767	100.0	18,247	100.0

Source: Economic Insights analysis.

2.4 Key Contract Parameters

Tables 2.2 and 2.3 present summary information on key contract variables for 2011 and 2014 respectively. These include the monthly charge, connection fee, contract term and the data transfer rate (or ‘capacity’) in megabits per second (Mbps) and the distance of the route over which the data is transferred. The latter two variables are the most important service characteristics.

Average charges on regulated routes were considerably lower than those on deregulated routes in 2011. The monthly average charge on regulated routes was \$1,219, compared to the average monthly charge on deregulated routes of \$1,898. This difference reflected both higher capacity and longer distances for typical services on deregulated routes. By 2014, the monthly charges on deregulated routes were similar to those on regulated routes, even though the average contracted capacity and distance for services on deregulated routes remained much higher than on regulated routes.

There was growth in the overall number of services in the datasets from 13,470 in 2011 to 18,247 in 2014. This is largely due to the inclusion of stand-alone tail-end services, and the absence of one provider in the 2011 sample. After adjusting for these two changes, the number of contracts decreased between these two periods at an annual rate of approximately 1.9 per cent. On the other hand, the average capacity of all contracts increased more than threefold from 43 Mbps in 2011 to 154 Mbps in 2014, and the average distance of a transmission service also increased by almost 50 per cent, from 90 km to 133 km.

Tables 2.2 and 2.3 also provide information on the distribution of values for each variable, including the mean, coefficient of variation, the minimum and maximum values and the 5th and 95th percentiles. There is a high degree of variation in annual charges and key terms and conditions between contracts, and the presence of extreme values is indicated by the fact that maximum values of variables are commonly far higher than the 95th percentile values.

The 2012 study found that there was considerable skewness in the distribution of most of the key variables, and for this reason they were transformed into logarithms, which reduced the skewness considerably. Figure 2.1 shows a histogram of the log monthly charge in 2014 for deregulated routes. Figures 2.2 to 2.4 show histograms for the log of capacity, log of distance and the log of the product of capacity and distance (in Mbps-km), in each case for 2014 for deregulated routes.

These charts indicate that a large proportion of contracts have similar capacity of around 2 Mbps. There are a large number of contracts with distances around 600-700 km, representing inter-capital routes. The log of the product of capacity and distance, Mbps-km, has a similar frequency distribution to the log monthly charge.

Table 2.2: Summary statistics, Contract parameters (2011 data)

	Obs	Missing	Mean	Coef. Var.	Min.	P(.05)	P(.95)	Max
<u>1. All routes</u>								
Monthly charge	13,470	0	1,425	3.49	█	245	4,650	█
Connection charge	2,309	11,161	6,618	21.85	█	0	10,000	█
Capacity (Mbps)	13,470	0	43	10.02	2	2	100	10,000
Distance (km)	13,470	0	90	3.56	0	1	594	3,611
Term (mths)	2,342	11,128	23	0.59	1	12	36	180
<u>2. Regulated routes</u>								
Monthly charge	9,375	0	1,219	3.90	█	245	3,473	█
Connection charge	739	8,636	3,243	6.38	█	0	10,000	█
Capacity (Mbps)	9,375	0	17	16.1	2	2	20	10,000
Distance (km)	9,375	0	67	3.48	0.5	3	358	3,413
Term (mths)	737	8,638	25	0.57	2	12	60	60
<u>3. Deregulated routes</u>								
Monthly charge	4,095	0	1,898	2.85	█	189	7,200	█
Connection charge	1,570	2,525	8,206	21.3	█	0	10,000	█
Capacity (Mbps)	4,095	0	102	6.42	2	2	280	10,000
Distance (km)	4,095	0	144	3.2	0	1	732	3,611
Term (mths)	1,605	2,490	22	0.59	1	12	36	180

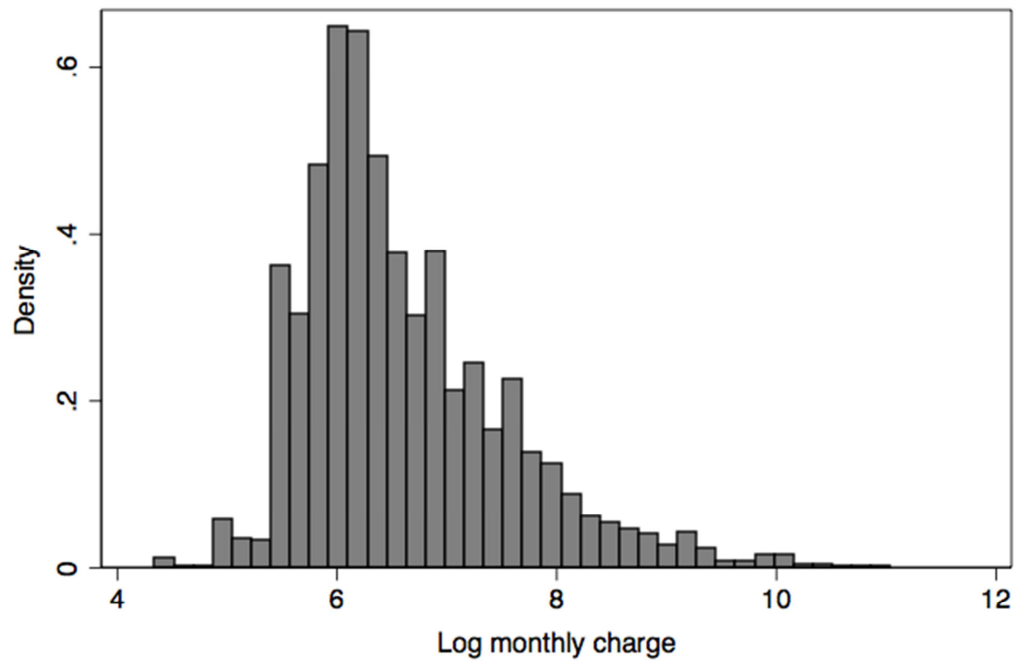
Source: Economic Insights analysis.

Table 2.3: Summary statistics, Contract parameters (2014 data)

	Obs	Missing	Mean	Coef. Var.	Min.	P(.05)	P(.95)	Max
<u>1. All routes</u>								
Monthly charge	18,247	0	1,386	2.49	■	243	4,609	■
Connection charge	18,247	0	2,013	21.54	■	0	3,300	■
Capacity (Mbps)	18,247	0	154	6.07	2	2	500	10,000
Distance (km)	18,247	0	133	3.24	0	0	728	3,618
Term (mths)	17,717	530	20	0.73	0	12	36	360
<u>2. Regulated routes</u>								
Monthly charge	11,480	0	1,399	2.67	■	243	4,671	■
Connection charge	11,480	0	1,549	22.36	■	0	3,300	■
Capacity (Mbps)	11,480	0	76	7.19	2	2	300	10,000
Distance (km)	11,480	0	107	3.44	0	0	600	3,608
Term (mths)	11,136	344	19	0.66	0	12	36	120
<u>3. Deregulated routes</u>								
Monthly charge	6,767	0	1,366	2.15	■	250	4,500	■
Connection charge	6,767	0	2,801	19.66	■	0	4,000	■
Capacity (Mbps)	6,767	0	288	4.71	2	2	1,000	10,000
Distance (km)	6,767	0	177	2.93	0	1	736	3,618
Term (mths)	6,581	186	23	0.79	0	12	36	360

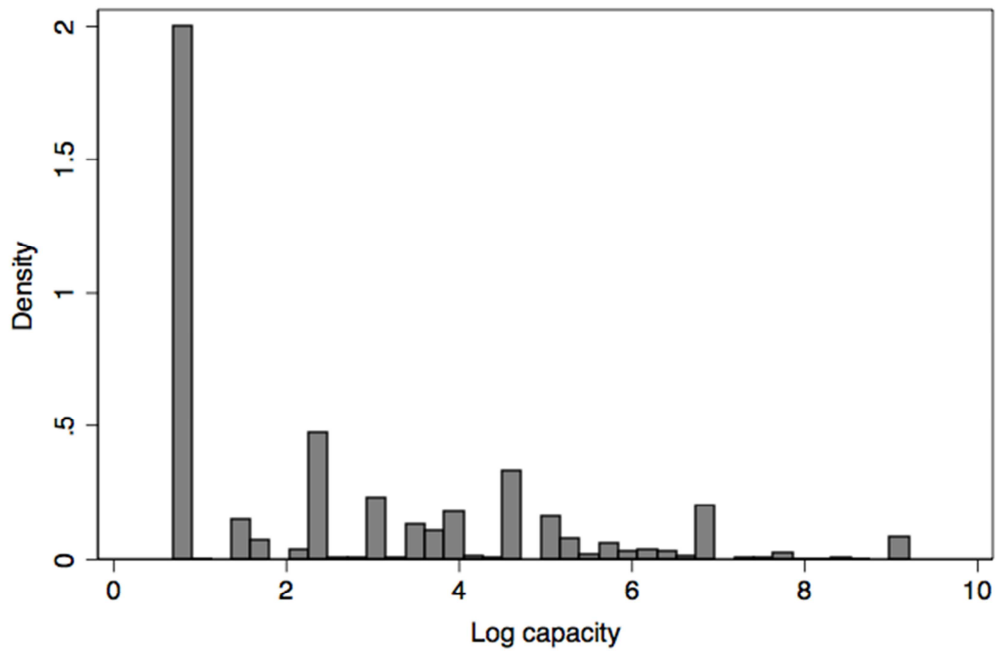
Source: Economic Insights analysis.

Figure 2.1: **Histogram: Log monthly charge (2014), deregulated routes**



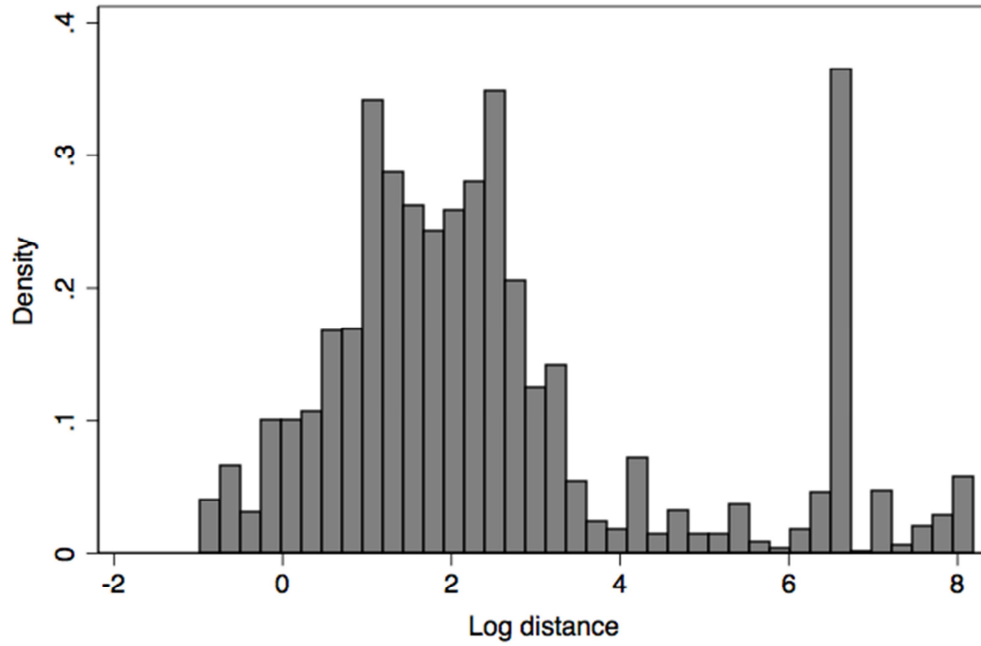
Source: Economic Insights analysis.

Figure 2.2: **Histogram: Log capacity (2014), deregulated routes**



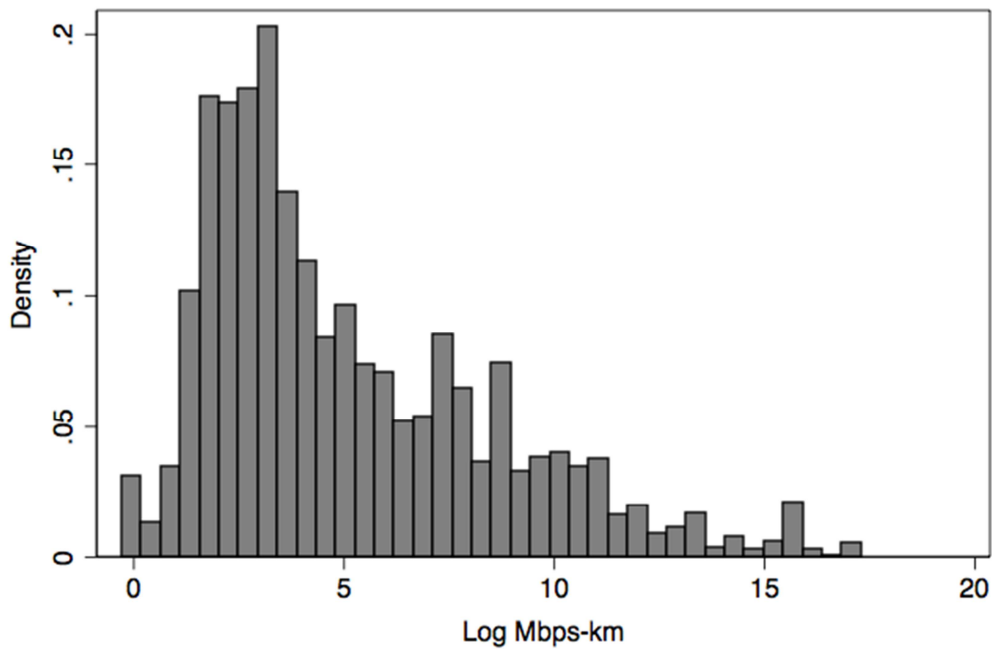
Source: Economic Insights analysis.

Figure 2.3: **Histogram: Log distance (2014), deregulated routes**



Source: Economic Insights analysis.

Figure 2.4: **Histogram: Log Mbps-km (2014), deregulated routes**



Source: Economic Insights analysis.

2.5 Quality of Service Indicators

There are three measures relating to quality of service in the dataset, namely: protection, interface-type and the ACCC's quality rating.

2.5.1 Protection

Protection refers to the existence of back-up facilities in the event of an interruption. In the 2011 dataset, the protection variable only indicates whether there is or is not any protection. The 2014 dataset has two different types of protection, electronic and geographic, but only 234 observations (1.3 per cent of all observations) have electronic protection, compared to 13,023 with geographic protection (71.4 per cent of all observations). We have defined a service as 'protected' if there is either electronic or geographic protection, and unprotected otherwise.

Table 2.4 shows a summary of the number of services with and without protection on deregulated and declared routes in the two periods. Two key observations may be made. Firstly, the proportionate coverage of protection has declined for both declared and deregulated routes between the two periods. In 2011, 88 per cent of all services (on deregulated and declared routes) had protection, and this declined to 72 per cent in 2014. Secondly, there is a much higher rate of protection on the regulated routes than on deregulated routes. In 2011, 95 per cent of services on declared routes had protection, compared with 74 per cent of services on deregulated routes. In 2014, 80 per cent of services on declared routes had protection compared with 58 per cent on deregulated routes. Both of these observations are partly associated with the increased share of the market supplied by 2nd and 3rd tier providers, especially on deregulated routes.

Table 2.4: **Protection by regulatory status, 2011 & 2014**

<i>Regulatory status</i>	2011			2014		
	Unprotected	Protected	Total	Unprotected	Protected	Total
No.						
Regulated	484	8,891	9,375	2,250	9,230	11,480
Deregulated	1,081	3,014	4,095	2,837	3,930	6,767
Total	1,565	11,905	13,470	5,087	13,160	18,247
Row%						
Regulated	5.2	94.8	100.0	19.6	80.4	100.0
Deregulated	26.4	73.6	100.0	41.9	58.1	100.0
Total	11.6	88.4	100.0	27.9	72.1	100.0

Source: Economic Insights analysis.

2.5.2 Interface

There are two interface types in the 2011 dataset, SDH (synchronous digital hierarchy) and Ethernet. In the 2014 dataset there are two additional interface categories, EoSDH (Ethernet

over SDH) and DWDM.⁶ Only 20 observations in the dataset have the DWDM interface type. We combined DWDM with Ethernet to form three classifications, namely SDH, Ethernet and EoSDH. Table 2.5 shows a summary of interface types in 2011 and 2014.

Overall, the share of Ethernet interfaces increased from 13 to 17 per cent over the two periods. Ethernet interfaces are used on a higher proportion of services on deregulated routes compared with declared routes. In 2014, 28 per cent of services on deregulated routes used Ethernet interfaces, compared with 11 per cent on regulated routes. Similarly, EoSDH interfaces are more prevalent on deregulated routes than on regulated routes. In 2014, 20 per cent of deregulated routes used EoSDH interfaces compared with 7 per cent for regulated routes.

Table 2.5: Interface type by regulatory status, 2011 & 2014

<i>Regulatory status</i>	2011			2014			
	Ethernet	SDH	Total	Ethernet*	EoSDH	SDH	Total
No.							
Regulated	536	8,839	9,375	1,262	812	9,406	11,480
Deregulated	1,252	2,843	4,095	1,862	1,375	3,530	6,767
Total	1,788	11,682	13,470	3,124	2,187	12,936	18,247
Row%							
Regulated	5.7	94.3	100.0	11.0	7.1	81.9	100.0
Deregulated	30.6	69.4	100.0	27.5	20.3	52.2	100.0
Total	13.3	86.7	100.0	17.1	12.0	70.9	100.0

Note: * includes DWDM.

Source: Economic Insights analysis.

2.5.3 Quality of service scores

The dataset includes a quality rating of each provider formulated by the ACCC, on a scale of 1 to 4, where 1 represents the highest standard and 4 the lowest. This is a subjective ordinal ranking, rather than a cardinal measure of quality. DAA suggested it was effectively a proxy for the provider, since in 2011 there were only seven providers, with the largest two, in the highest quality categories, accounting for [REDACTED] of the services on the deregulated routes, and with the smaller providers grouped into the lowest two quality categories. This situation has changed to some extent in the 2014 dataset, because:

- the number of providers increased from seven to nine;

- [REDACTED]

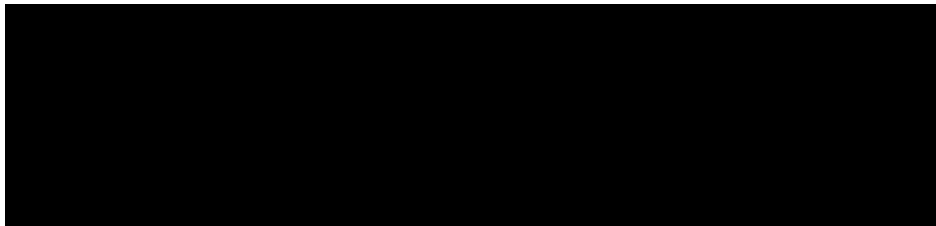

- [REDACTED]

One possible concern with this measure is that larger providers tend to have higher quality scores, so that it may be correlated with other factors relevant to pricing. It is also correlated

⁶ ACCC staff advised that the two new categories were treated as part of SDH in 2011-12.

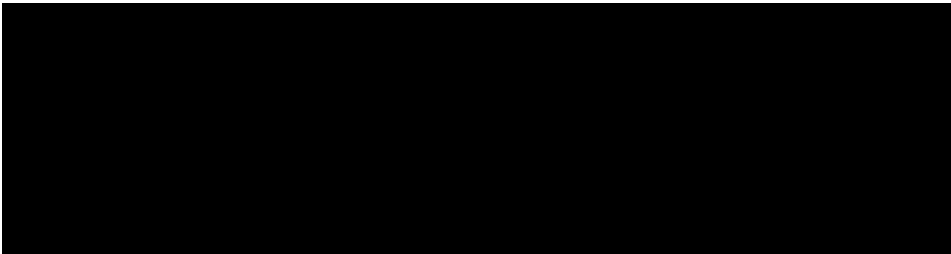
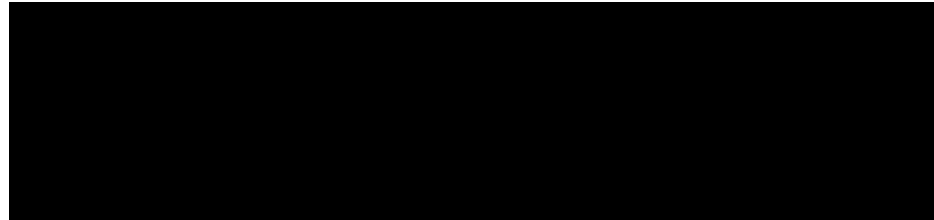
with other indicators of quality of service, including protection and interface type. This is shown in tables 2.6 and 2.7. Table 2.6 cross-tabulates protection with the ACCC’s quality rating, and table 2.7 cross-tabulates interface type with the ACCC’s quality rating, in both cases for deregulated routes only.

Table 2.6: Protection by provider quality class, 2011 & 2014, deregulated routes

<i>Provider quality class</i>	2011			2014		
	Unprotected	Protected	Total	Unprotected	Protected	Total
No.						
1						
2						
3						
4						
Total						
Row%						
1						
2						
3						
4						
Total						

Source: Economic Insights analysis.

Table 2.7: Interface type by provider quality class, 2011 & 2014, deregulated routes

<i>Provider quality class</i>	2011			2014			
	Ethernet	SDH	Total	Ethernet	EoSdh	SDH	Total
No.							
1							
2							
3							
4							
Total							
Row%							
1							
2							
3							
4							
Total							

Source: Economic Insights analysis.

2.6 Indicators of market size and competition

The datasets contain a number of variables that appear to be related to market size and competition on a specific route or in the ESAs joined by a specific route. Some variables of this kind that were included in the 2011 dataset were not available in the 2014 dataset, but the omitted variables had little correlation with monthly charges. Table 2.8 presents summary information for selected variables of this kind.

Some of the variables of this kind are indicators of the number of households or end-users in the ESAs at the A and B ends of a route, which are related to the size of the markets in those ESAs. They include:

- the number of relevant addresses in the Geocoded National Address File (GNAF)⁷ (not available in 2014)
- population and population density (not available in 2014)
- the number of services in operation (SIOs)
- the average area of those ESAs in km²
- the density of SIOs per km².

These measures primarily relate to the size or density of overall telecommunications demand in the ESAs at the ends of a route, and do not appear to *directly* relate to the demand for telecommunications transmission services that a single provider faces on a particular route. This is partly because the ESAs may have more routes originating or terminating at them, and partly because the indicators relate to retail consumers. To the extent that they influence DTCS costs and prices, their influence would be indirect and conditioned by other factors. However, these measures were mostly found to have very little correlation with monthly charges and there are insufficient economic grounds for considering them as determinants of DTCS provider costs. For these reasons they are not included in the econometric modelling.

The amount of competition in supplying the end-user markets in the end-point ESAs with traditional voice services, broadband internet and VoIP services may be indicated by the variable:

- Average # of ULLS/LSS access seekers at the relevant ESAs.

This is described by the ACCC as an indicator of the derived demand for transmission services. However, there are direct measures of the demand for transmission services in the dataset, including route throughput. In the workshop paper this variable was tested as a competition variable and found to be less useful than the other variables included in this study as indicators of competition. In the present context it is likely to be unreliable as an indicator of competition in DTCS services because some of the access seekers to the copper network to provide services to end-users may not be competitors in the transmission market and vice versa.

⁷ A national database of addresses sourced from Australian governments, the Australian Electoral Commission and Australia Post.

Other variables related *more directly* to market size and competition on transmission routes, and which are included in both the 2011 and 2014 data, include:

- (1) Average # of telecommunications providers with a presence at the relevant ESAs
- (2) Route throughput (Mbps)
- (3) Average throughput at the end-point ESAs (Mbps)
- (4) DTCS service provider throughput on relevant route (Mbps)
- (5) # of DTCS providers on the specific route (i.e. number of DTCS providers with contracts on the specific route — see Annex A for more detail)
- (6) # of DTCS services on the specific route (i.e., number of records in the dataset on each route).

A number of these variables are based on aggregating the number of contracts or contract capacity data by route or by ESA to derive the number of services, or throughput measures, or the number of DTCS providers on specific routes or the end-point ESAs. These variables may potentially explain some differences in costs of the various providers on the different routes and are included in the econometric analysis.

Table 2.8: Summary statistics, selected contextual variables, deregulated routes

	Obs	Missing	Mean	Coef. Var.	Min.	P(.05)	P(.95)	Max
A. 2011 data								
Average # ULLS/LSS access seekers	4,095	0	8.5	0.23	0.5	4.5	11	12
Average # telco. providers at ESAs	4,095	0	5.5	0.23	2.5	3.5	7.5	8.5
Route throughput (Mbps)	4,095	0	1,564	2.59	2	2	10,150	35,605
ESAs throughput (Mbps)	4,095	0	58,434	0.82	26	921	145,444	200,797
Route throughput of DTCS provider	4,095	0	579	3.92	2	2	2,712	35,605
Average # of SIOs	4,095	0	18,504	0.24	7,122	12,368	16,453	31,360
Average ESA size (km ²)	4,095	0	13.6	1.56	0.5	1.6	40	510.4
SIO density (per km ²)	4,095	0	3,383	1.23	28	503	9,295	28,796
B. 2014 data								
Average # ULLS/LSS access seekers	11,480	0	4.7	0.49	0	1	9	11
Average # telco. providers at ESAs	11,480	0	3.7	0.35	1	2	6	8
# DTCS providers on route	11,480	0	1.5	0.64	1	1	4	6
# DTCS services on route	11,480	0	15.8	1.35	1	1	59	114
Route throughput (Mbps)	11,480	0	1,162.1	3.94	2	2	4,794	42,607
Route throughput of DTCS provider	11,480	0	557.4	4.91	2	2	2,055	42,004
ESAs throughput (Mbps)	11,480	0	87,321.6	1.67	4	240	410,268	821,179
Average # of SIOs at ESAs	11,480	0	12,669.1	0.42	110	5,015	22,891	31,957
Average ESA size (km ²)	11,480	0	204.9	11.02	1	6	561	91,097
SIO density (per km ²)	11,480	0	828.0	2.54	0	13	2,252	27,760

Source: Economic Insights analysis.

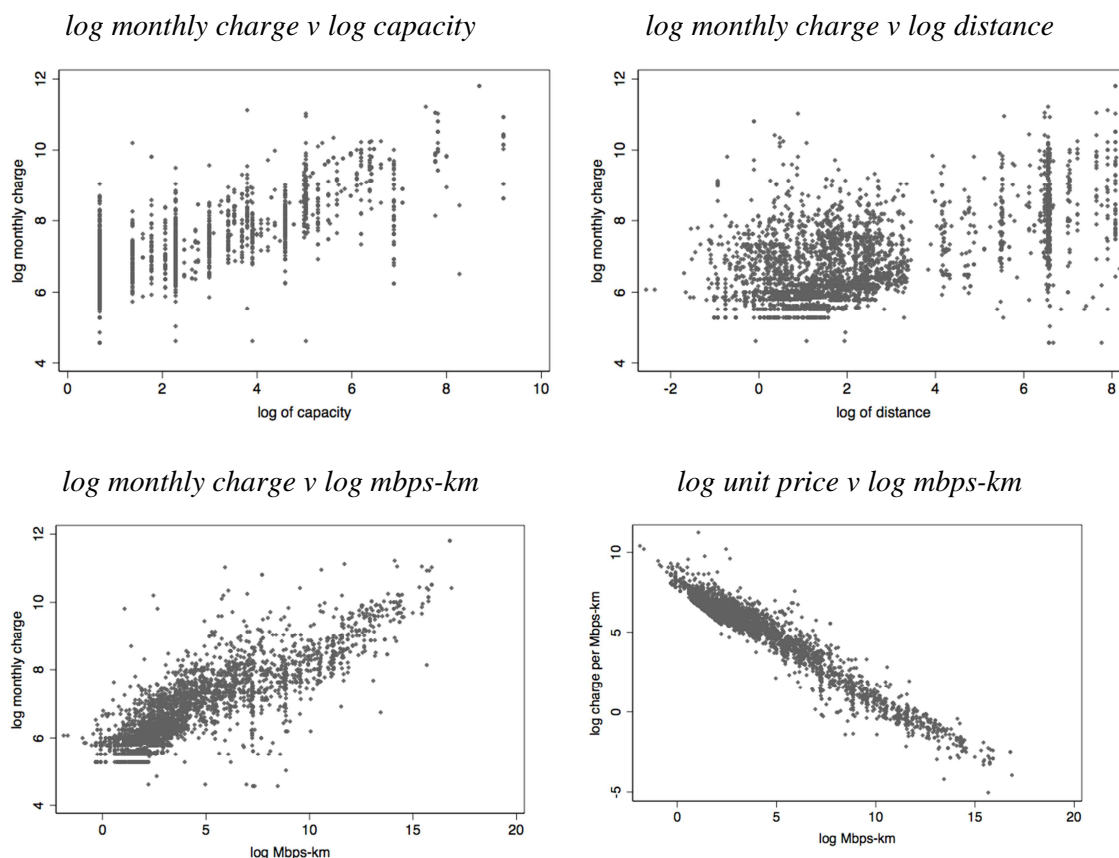
3 EXPLORATORY DATA ANALYSIS

The 2012 study found that the key predictors for charges were the distance and capacity attributes of the service. Other relevant predictors were the route type, quality of service, protection, and interactions between some of these factors. Demand variables, discussed in the previous section, did not feature in the final model. This section presents exploratory data material relating to the 2011 and 2014 data. The exploratory analysis is largely in the form of data plots, tables and discussion.

3.1 Capacity and Distance

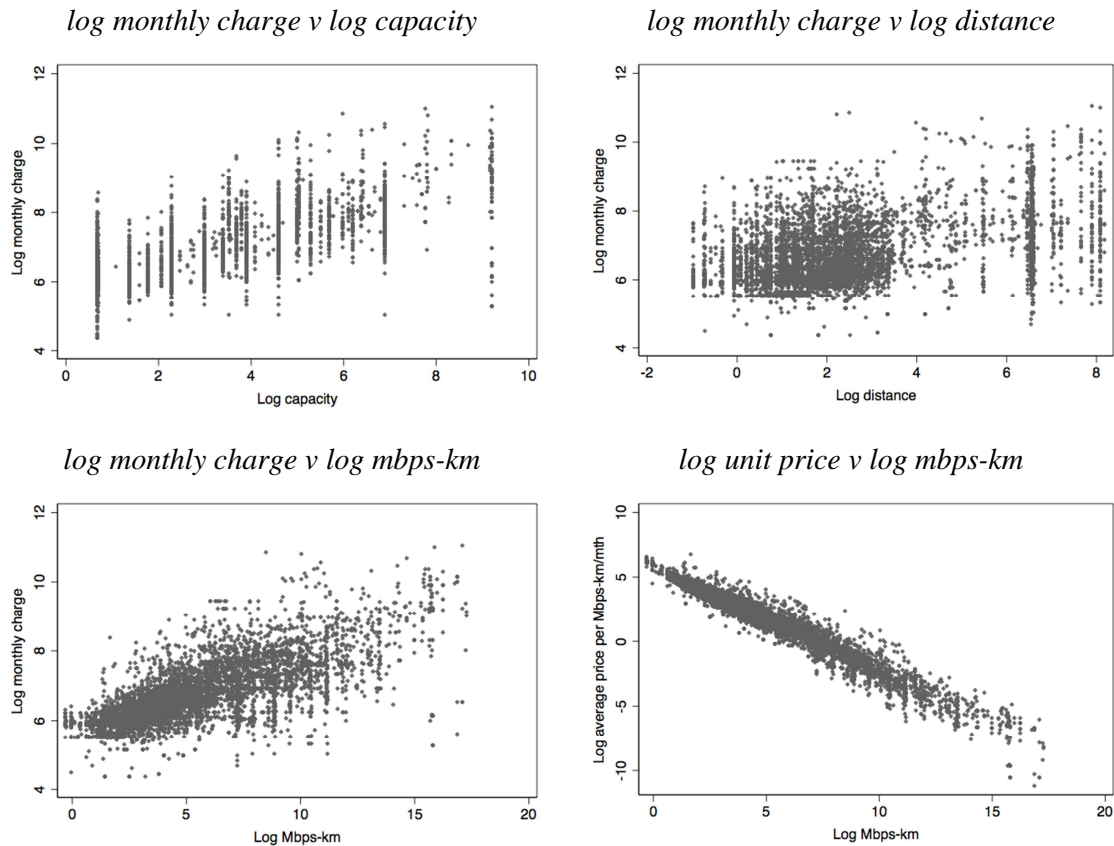
Scatter diagrams are useful tools for visualising the relationships between variables. The bivariate relationships between the log of annual charges and the log values of capacity and distance in both the 2011 and 2014 datasets are presented in Figures 3.1 and 3.2, each of which includes several scatter plots.

Figure 3.1: **Scatter diagrams: log monthly charge v output measures (2011 data)**



Source: Economic Insights analysis.

 Figure 3.2: **Scatter diagrams, log monthly charge v output measures (2014 data)**



Source: Economic Insights analysis.

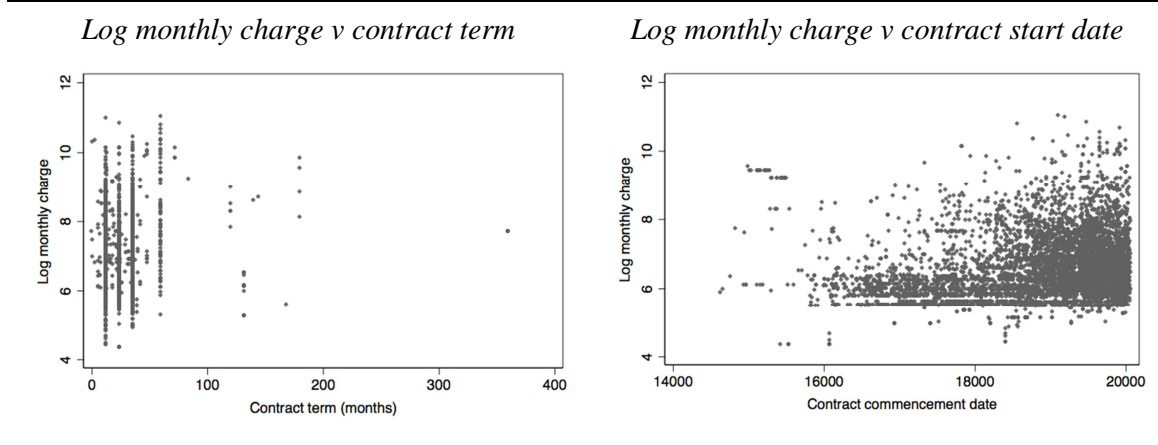
One question is whether there is value in developing a single measure of output that combines the capacity and distance dimensions, such as the Mbps-km variable previously discussed. The monthly charge paid by a user is equivalent to a total cost or revenue concept, and when divided by a single measure of output, the result is an average cost or unit price measure.

The lower right-hand quadrants of Figures 3.1 (2011) and 3.2 (2014) plot the log of average cost or unit price (i.e., the monthly charge for a service divided by the Mbps-km) against the log Mbps-km. In this form, it appears to be a comparatively well-defined negative relationship, with an increasing slope at higher values of Mbps-km.

3.2 Contract start date and term

Figure 3.3 provides scatter plots for the relationship between the log monthly charge and the contract term and the contract start date. There is no clear relationship between the log monthly charge and the contract term but there is a positive relationship between the log monthly charge and the log of the contract start date, given the pattern of observations in the bottom right hand corner of the scatter plot.

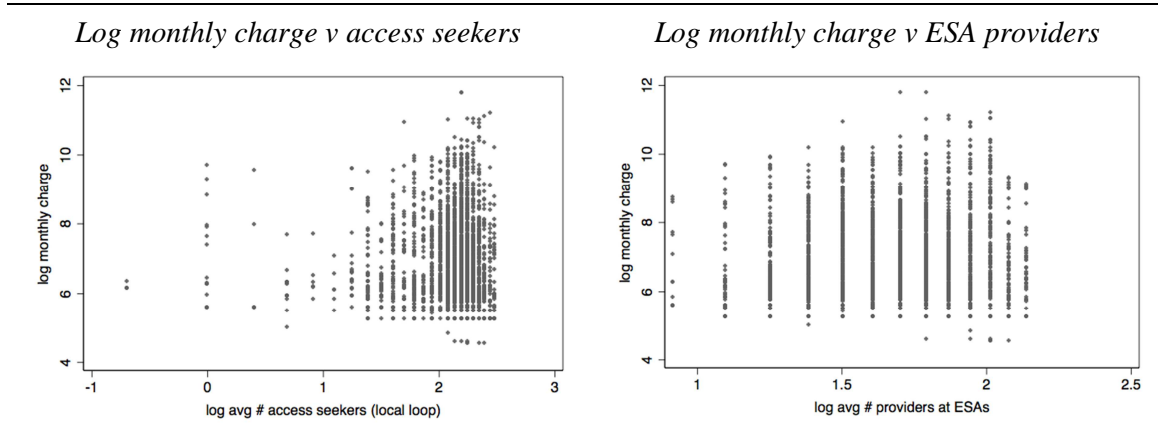
Figure 3.3: **Scatter diagrams: log monthly charge v other contract data (2014 data)**



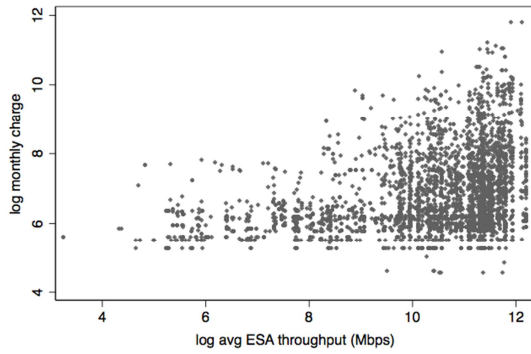
3.3 Conditioning variables

Figure 3.4 show scatter plots indicating the bivariate relationships between the log of annual charge and several demand-related variables. Several of demand-related variables in the dataset are alternative measures of the same variable (e.g. SIOs), and in these cases only one of the variables is presented. Only variables that are present in both the 2011 and 2014 data are shown. Those shown include: the log number of ULLS/LSS access seekers; the number of ‘suppliers’ or providers with a presence at the relevant ESAs; average throughput of the relevant ESAs; average ESA size in km²; the throughput of the specific provider on the specific route; the overall throughput of all providers on the relevant route; the average number of SIOs at the A-end and B-end ESAs; and the average SIO density per km². Several of these relationships exhibit only moderate correlation, and in some cases it appears that a correlation emerges only after a threshold value is reached.

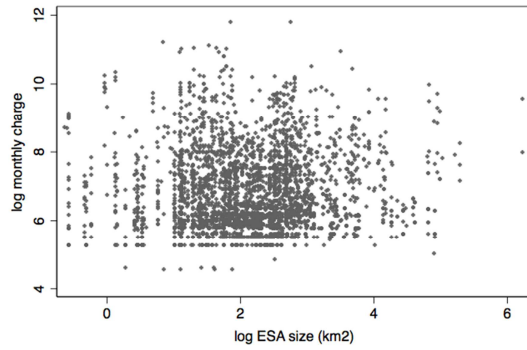
Figure 3.4: **Scatter diagrams: log monthly charge v contextual variables (2011 data)**



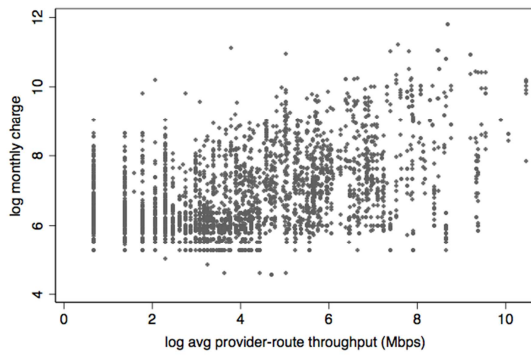
Log monthly charge v log ESA throughput



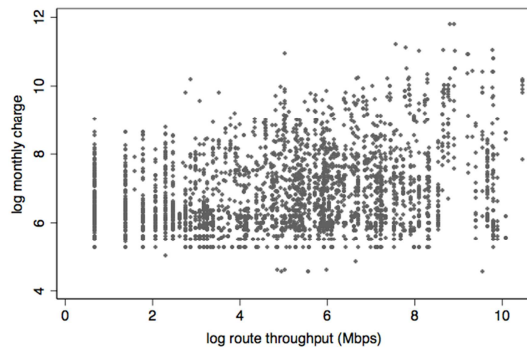
Log monthly charge v log ESA size



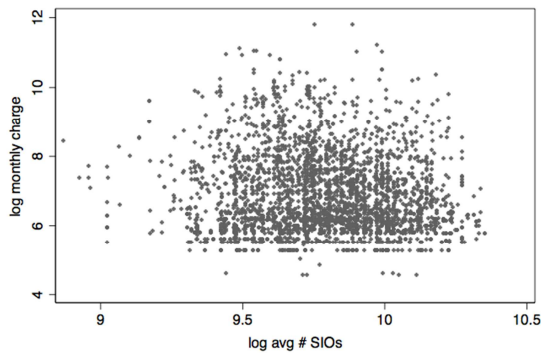
Log monthly charge v log provider-route throughput



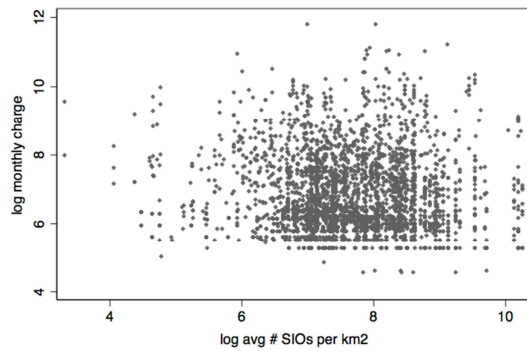
Log monthly charge v log route throughput



Log monthly charge v avg. # SIOs

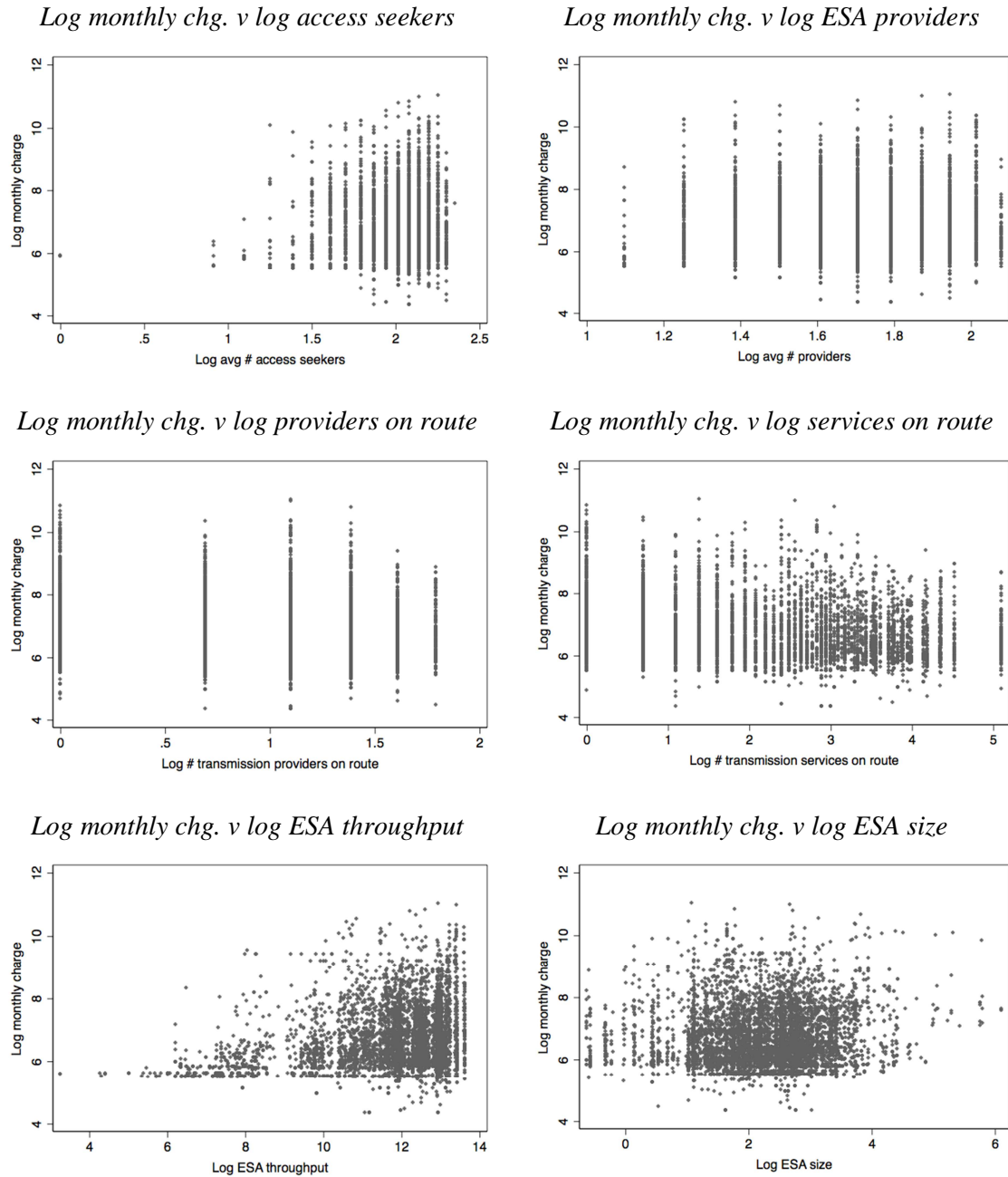


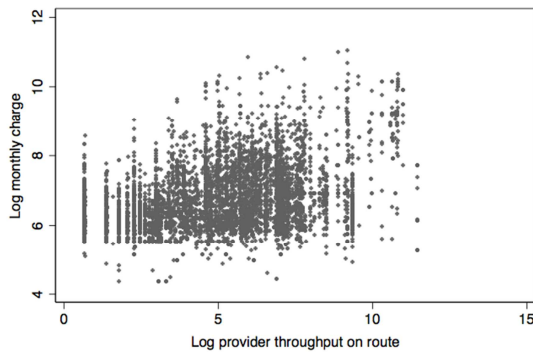
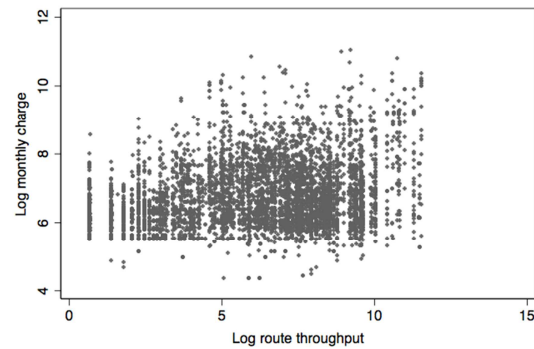
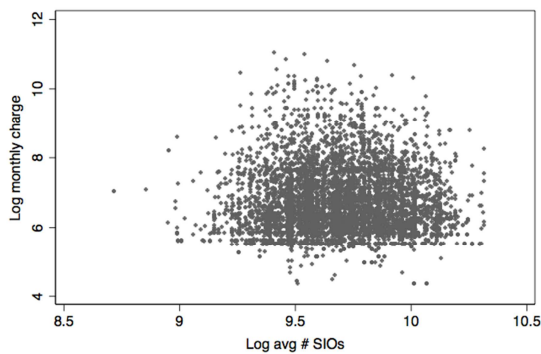
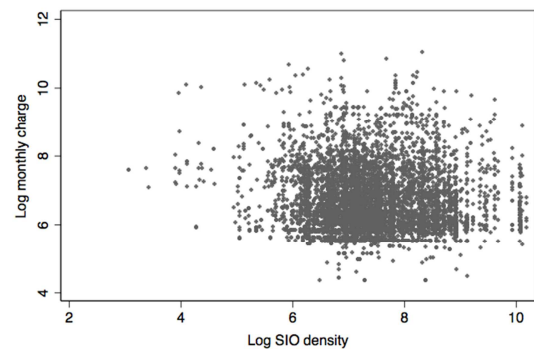
Log monthly charge v SIO density



The corresponding relationships in the 2014 data are shown in Figure 3.5. Some of these variables may be related to economies of scale at some exchanges.

Figure 3.5: **Scatter diagrams, log monthly charge v supply-demand factors (2014 data)**



Log monthly chg. v log provider-route t'put*Log monthly chg. v log route t'put**Log monthly chg. v log avg. # SIOs**Log monthly chg. v log SIO density*

The foregoing charts suggest that there appears to be a linear relationship between the log monthly charge and log route throughput and log provider-route throughput. These two variables, log route throughput and log provider-route throughput, are strongly correlated with each other because on average there are few providers on each route. It is likely to be problematic to use both variables in an econometric analysis.

Tables 3.1 and 3.2 present correlation coefficients for the continuous variables in the 2011 and 2014 datasets. Some of the variables in the 2011 dataset have been excluded because they are not present in the 2014 dataset, such as those based on GNAFs (geocoded national address files) and population data. Further, both datasets include four different measures of the number of SIOs that not only have very high correlation with each other, but also very similar correlation coefficients with the other variables in the datasets, so only one of these measures is shown in each table.

Table 3.1: Correlation coefficients for log values of continuous variables, deregulated routes (2011 data)

<i>Log of variable</i>	Annual Charge	Unit Price	Connect. charge	Capacity	Distance	Term	Avg access seekers	Avg sup-pliers	Avg # SIOs	Avg ESA size	SIO density	Route th'put	ESA th'put	Provider route-th'put
Monthly Charge (\$)	1.0000													
Unit Price (\$/ Mbps-km)	-0.7039*	1.0000												
Connection charge (\$)	0.5575*	-0.3345*	1.0000											
Capacity (mbps)	0.8235*	-0.6909*	0.4753*	1.0000										
Distance (km)	0.5982*	-0.8821*	0.2063*	0.3380*	1.0000									
Term (mths)	0.1645*	-0.2270*	0.1149	0.2737*	0.1055*	1.0000								
Avg # access seekers	0.0200	0.0535	0.0214	0.1131*	-0.1459*	0.0473	1.0000							
Avg # ESA providers	0.0439	0.0925*	0.0014	0.1467*	-0.2053*	-0.0887	0.6122*	1.0000						
Avg # SIOs	-0.1025*	0.0874*	-0.0125	-0.0750*	-0.0863*	-0.0658	0.3157*	0.0727*	1.0000					
Avg size of ESA (km ²)	0.0619*	-0.2351*	0.0709	-0.0905*	0.3705*	0.0572	-0.5229*	-0.6833*	0.0704*	1.0000				
SIO density	-0.0867*	0.2537*	-0.0726	0.0705*	-0.3869*	-0.0719	0.5945*	0.6919*	0.1806*	-0.9684*	1.0000			
Route t'put (mbit/s)	0.2315*	-0.0716*	0.1297*	0.4187*	-0.1589*	0.0412	0.3375*	0.4162*	0.0197	-0.3684*	0.3682*	1.0000		
ESA t'put (mbit/s)	0.2941*	-0.1763*	-0.0072	0.3122*	0.0818*	-0.0150	0.3909*	0.4325*	-0.0045	-0.3475*	0.3416*	0.5970*	1.0000	
Provider route-t'put	0.3996*	-0.2170*	0.2083*	0.6026*	-0.0676*	0.0868	0.2450*	0.3413*	0.0237	-0.2802*	0.2822*	0.7963*	0.4311*	1.0000
Mbps-km	0.8577*	-0.9689*	0.4343*	0.7869*	0.8468*	0.2274*	-0.0317	-0.0517	-0.0990*	0.1918*	-0.2138*	0.1325*	0.2301*	0.2963*

Note: * Significant at the 5% level or better.

Table 3.2: Correlation coefficients for log values of continuous variables, deregulated routes (2014 data)

<i>Log of variable</i>	<i>Monthly Charge</i>	<i>Connect. charge</i>	<i>Avg. Price</i>	<i>Capacity</i>	<i>Distance</i>	<i>Mbps-km</i>	<i>Contract start date</i>	<i>Contract term</i>	<i>Avg # access seekers</i>
Monthly Charge (\$)	1								
Connection charge (\$)	0.5731*	1							
Avg. Price (\$/ Mbps-km)	-0.5645*	-0.3011*	1						
Capacity (mbps)	0.7568*	0.4048*	-0.7325*	1					
Distance (km)	0.4110*	0.2146*	-0.8253*	0.2746*	1				
Mbps-km	0.7357*	0.4129*	-0.9744*	0.8073*	0.7892*	1			
Contract start date	0.2120*	-0.0245	-0.2976*	0.3556*	0.1220*	0.3020*	1		
Contract term (mths)	0.2267*	0.2230*	-0.2308*	0.3034*	0.0937*	0.2512*	0.0707*	1	
Avg # access seekers	0.0967*	-0.2108*	-0.0880*	0.1650*	-0.0111	0.0986*	0.0999*	0.0746*	1
Avg # ESA providers	0.1281*	-0.1682*	-0.0400	0.1600*	-0.0562*	0.0677*	0.1242*	0.0324	0.6722*
# DTCS providers	0.036	-0.1826*	0.0679*	0.1791*	-0.2612*	-0.0459*	0.1942*	0.0499*	0.4963*
# DTCS providers (ex. top 4)	-0.0281	-0.0536	0.1179*	0.0341	-0.2008*	-0.1068*	0.0554	-0.0571	0.1297*
DTCS services	-0.1880*	-0.2452*	0.3011*	-0.0213	-0.4639*	-0.2983*	0.0916*	-0.0212	0.3302*
Route th'put (mbit/s)	0.2589*	0.0334	-0.2242*	0.4691*	-0.0735*	0.2545*	0.2861*	0.1658*	0.4514*
Provider route-th'put	0.3782*	0.1355*	-0.3358*	0.6193*	-0.0277	0.3786*	0.2966*	0.1955*	0.3212*
ESA throughput (mbit/s)	0.2830*	-0.0279	-0.2788*	0.3296*	0.1553*	0.3059*	0.2611*	0.1463*	0.5675*
Avg # SIOs	-0.0147	-0.0884*	-0.0093	0.0212	-0.0162	0.0036	-0.0185	-0.0166	0.2726*
Avg size of ESA (km2)	0.0357	0.1941*	-0.1487*	-0.0496*	0.2663*	0.1318*	-0.0371	0.0489*	-0.4831*
SIO density	-0.0393	-0.2154*	0.1463*	0.0548*	-0.2701*	-0.1308*	0.0325	-0.0529*	0.5506*

Table 3.2: (cont.)

<i>Log of variable</i>	<i>Avg # ESA providers</i>	<i># DTCS providers</i>	<i># DTCS providers (ex. top 4)</i>	<i># DTCS services</i>	<i>Route t'put</i>	<i>Provider route-t'put</i>	<i>ESA t'put</i>	<i>Avg # SIOs</i>	<i>Avg size of ESA</i>
Avg # ESA providers	1								
# DTCS providers	0.4595*	1							
# DTCS providers (ex. top 4)	0.1812*	0.4579*	1						
# DTCS services	0.3378*	0.7839*	0.3996*	1					
Route th'put (mbit/s)	0.4418*	0.7919*	0.3203*	0.6953*	1				
Provider route-th'put	0.3297*	0.5374*	0.2518*	0.5529*	0.8110*	1			
ESA throughput (mbit/s)	0.5266*	0.6150*	0.2500*	0.3613*	0.6226*	0.4779*	1		
Avg # SIOs	-0.0155	0.0939*	-0.0389	0.0344	0.0529*	0.0054	0.0298	1	
Avg size of ESA (km2)	-0.5913*	-0.3477*	-0.1174*	-0.3146*	-0.3185*	-0.2232*	-0.3086*	0.1208*	1
SIO density	0.5869*	0.3708*	0.1069*	0.3229*	0.3314*	0.2244*	0.3158*	0.1283*	-0.9690*

Note: * Significant at the 5% level or better.

3.4 Categorical Variables

Measures of bivariate association between categorical variables are provided by Cramer's V statistic, which has a value of zero for no association and 1 for perfect association. Values greater than 0.6 represent strong association. These are presented in Table 3.3. The A and B-end states are strongly associated because a high proportion of services are provided within

one state. The only other strong associations are between the provider on the one hand, and the QOS, interface-type and protection of the service on the other.

The association between a continuous variable and a categorical variable can be measured by regressing the continuous variable against indicator variables for each value of the categorical variable, and calculating the square root of the resulting R^2 . These statistics are shown in Table 3.4. The monthly charge has only moderate bivariate association with the categorical variables.

Table 3.3: Cramer's V for categorical variables, 2015 data, deregulated routes

<i>Log of variable</i>	<i>Provider</i>	<i>ACCC</i>	<i>QOS</i>	<i>Interface</i>	<i>Protection</i>	<i>A-end</i>	<i>B-end</i>	<i>NBN</i>
		<i>route</i>		<i>type</i>		<i>State</i>	<i>State</i>	<i>POIs</i>
		<i>categories</i>						
Provider	1							
ACCC route categories	0.2529	1						
QOS	1.0000	0.2028	1					
Interface type	0.6077	0.1483	0.5754	1				
Protection	0.5969	0.1348	0.4133	0.0742	1			
A-end State	0.3449	0.2688	0.3165	0.1974	0.2025	1		
B-end State	0.3368	0.2459	0.3190	0.2023	0.2052	0.7151	1	
NBN POIs	0.1552	0.0707	0.1051	0.0836	0.0033	0.2826	0.2839	1

Table 3.4: Correlation of log values of continuous variables against categorical variables (2015 data, deregulated routes)*

<i>Log of variable</i>	<i>Provider</i>	<i>ACCC</i>	<i>QOS</i>	<i>Interface</i>	<i>Protection</i>	<i>A-end</i>	<i>B-end</i>	<i>NBN</i>
		<i>route</i>		<i>type</i>		<i>State</i>	<i>State</i>	<i>POIs</i>
		<i>categories</i>						
Monthly Charge (\$)	0.4539	0.3564	0.3137	0.3762	0.1851	0.0994	0.0919	0.0570
Connection charge (\$)	0.5059	0.2882	0.2139	0.2853	0.0999	0.1952	0.1755	0.0291
Avg. Price (\$/ Mbps-km)	0.5277	0.7370	0.3918	0.3892	0.1721	0.1415	0.1093	0.1137
Capacity (mbps)	0.5943	0.2758	0.4544	0.5233	0.1774	0.0890	0.1076	0.0631
Distance (km)	0.3173	0.8615	0.2341	0.1529	0.1276	0.1962	0.1206	0.1276
Mbps-km	0.5424	0.7016	0.4063	0.4130	0.1916	0.1354	0.1033	0.1063
Contract start date	0.3648	0.1461	0.3511	0.4884	0.0931	0.1072	0.1150	0.0348
Contract term (mths)	0.3288	0.0568	0.1740	0.0937	0.1327	0.1198	0.1361	0.0516
Avg # access seekers	0.3774	0.4236	0.3759	0.2625	0.1782	0.3147	0.3240	0.0303
Avg # ESA providers	0.3897	0.2438	0.3803	0.2457	0.1471	0.1896	0.1908	0.1151
# DTCS providers	0.4448	0.2621	0.4418	0.3396	0.1649	0.2553	0.2567	0.0946
# DTCS providers (ex. top 4)	0.1369	0.1446	0.0950	0.0596	0.0193	0.1815	0.1507	0.2525
DTCS services	0.2788	0.3171	0.2574	0.1699	0.0775	0.2923	0.2826	0.0086
Route th'put (mbit/s)	0.5125	0.1748	0.4865	0.3983	0.2085	0.2253	0.2375	0.0590
Provider route-th'put	0.5783	0.1255	0.5259	0.4156	0.2779	0.2303	0.2380	0.0136
ESA throughput (mbit/s)	0.5199	0.2946	0.5177	0.3974	0.2461	0.2358	0.2514	0.1318
Avg # SIOs	0.1523	0.1277	0.1022	0.0774	0.0142	0.3510	0.3525	0.3763
Avg size of ESA (km2)	0.3410	0.3441	0.2821	0.1490	0.0696	0.2037	0.1968	0.1791
SIO density	0.3530	0.3707	0.2968	0.1613	0.0660	0.2445	0.2433	0.0852

* calculated as the square root of the R^2 of the regression of the continuous variable against the indicator variables for the categories of the categorical variables.

4 REVIEW OF 2012 DTCS ECONOMETRIC MODEL

This chapter discusses suitability for the present purpose of the econometric model used in the 2012 DTCS FAD, which was developed by Data Analysis Australia (DAA 2012). The 2012 econometric model (referred to as ‘the DAA 2012 model’) is examined in detail in Annex C using both the 2011 and 2014 datasets. This section summarizes the findings.

4.1 The 2012 econometric model

The form of the 2012 model is linear in the logarithms of capacity and distance, the route classification, the ACCC’s quality of service classification of providers (*qos*), and whether the service has protection. There are also interactions between the route classification and the ACCC’s quality of service categories, and between route classification and capacity. Variables relating to market size or competition were not included in the model. The model is summarised in the following equation:

$$(4.1) \quad R = \alpha_0 + \alpha_1 C + \alpha_2 D + \alpha_3 \text{protection} + \beta \cdot i.\text{route} + \varphi \cdot i.\text{qos} + \gamma \cdot (i.\text{route} \cdot i.\text{qos}) + \delta \cdot (i.\text{route} \cdot C)$$

where: *R* is the log annual charge in \$, *C* is log capacity, *D* is log distance, *protection* is an indicator variable indicating whether there is a back-up service, *i.route* is a set of indicator variables representing the route categories, and *i.qos* is a set of indicator variables representing ACCC’s quality of service categories.

The DAA 2012 model is replicated in Annex C using the 2011 dataset, and a range of diagnostic tests are presented. Also presented are the results of estimating the model with some changes to the specification to test the scope for improvement to the model. The changes that were tested include:

- adding higher-order terms to better address nonlinearities in the relationship between costs and outputs
- including additional conditioning variables which affect supplier costs
- allowing for unobserved route-specific effects on costs, by using a random effects specification rather than ordinary least squares (OLS).

The details of these estimated models are shown in Annex C. The results indicated that there is scope to make improvements to the specification in each of the three ways described. The desirability of adding conditioning variables depends on the balance to be struck between simplicity and statistical significance.

4.2 Diagnostic tests

Annex C presents a range of diagnostic tests relating to the DAA model and the alternative models estimating with the 2011 data sample. Most of the tests we have undertaken were not reported in the DAA study. The diagnostic tests applied to the DAA 2012 model are of two main types. Firstly, there are those that relate to the residuals, including tests of:

- whether the residuals are normally distributed (primarily needed only for hypothesis tests to be valid and also less relevant where asymptotic tests can be used for large samples)
- the ‘influence’ of individual observations, including outliers and observations that exert undue influence on the coefficients
- homoscedasticity (or constancy of variance) of the residuals.

Secondly, there are tests relating to the specification of the regression model include tests of:

- high multicollinearity between predictors (which may inflate the estimated variances, affecting the sign and magnitude of the coefficients)
- misspecification in terms of linearity of the functional relationship between the predictors and the dependent variable, the likelihood of omitted variables and the appropriateness of the dependent variable specification.

With regard to the DAA 2012 model, the following findings relate to the residuals. More detail is available in Annex C.

- *Normality of Residuals:* The formal statistical tests strongly rejected the null hypothesis of normality of the residuals. The distribution of residuals has fatter tails than the normal distribution. Normality of the distribution of residuals is not an essential requirement for unbiased estimates. It is necessary for valid hypothesis testing in small samples, but the sample size in this study is sufficiently large that non-normality of the residuals is not likely to be an issue of concern.
- *Homoscedasticity:* An important assumption of ordinary least squares (OLS) regression is the homogeneity of variance of the residuals. If the variance of the residuals is non-constant then there is heteroscedasticity; two implications are: (a) the relationship between the variables may be nonlinear, rather than the assumed linear relationship; and (b) conventional standard error estimates for the coefficients will be biased (making significance testing and inference unreliable), although White’s robust standard errors can be used in these circumstances. The formal statistical tests indicate that heteroscedasticity is present in the DAA model and the alternative models tested in Annex C.
- *Observations with Undue Influence:* A model may lack robustness if there are a small number of overly influential data points, which may cast doubt on inferences based on the model, or affect its performance in making out-of-sample predictions. The *influence* of an observation is the combined effect of being an *outlier* (where the residual term from the regression is large in absolute value) and having high *leverage* (where a predictor takes an extreme value relative to its mean). Using standard tests 60 observations (or 1.5 per cent of the sample) were identified as outliers and 20 were identified as severe outliers. A much greater number of observations had a high degree of leverage. We estimate that highly influential observations represent 3.4 per cent of the sample.

The DAA report made the following comment on the influence of data points:

Although there are several data points with large residual values, they do not have a large influence on the model, as indicated by their small leverage values. There are several other data points that have relatively high leverage. These data points, however, have small residuals indicating that the model fits these points well.

(DAA,2012 p.9)

While we accept that observations with high influence (i.e. are both outliers and have high leverage) are of most concern, other extreme outliers and observations with high leverage are also relevant because of their large effects on the estimates.

The tests reported in Annex C indicate that neither the DAA model nor the alternatives shown there, satisfied the tests relating to residuals. However, two points need to be emphasised. Firstly, it is not necessarily crucial that these tests be satisfied in large samples because the least squares estimator is unbiased and efficient. That is, in large samples the estimates should become more accurate if the model is correctly specified. A key problem is to correctly specify the model in the situation where extreme values are present. Secondly, these issues relate to characteristics of the data and are not easily resolved by altering the model specification, although alternative regression methods are available that give less weight to outliers (see chapter 5). Although these methods can assist to develop a suitable specification, none of the models discussed in this report resolve all issues relating to residuals.

The following findings relate to the specification of the DAA 2012 model:

- *Multicollinearity*: Multicollinearity can become a problem if there is close correlation between predictors, such that a substantial part of the variation of one of the predictors could be explained by a linear function of the other predictors. When there is a high degree of multicollinearity the coefficient estimates may be poorly identified with large variance and some coefficient estimates may be sensitive to small changes in the data. Multicollinearity is a sampling problem, in the sense that a larger and richer dataset may enable the poorly identified effects to be better identified.

One method of detecting high multicollinearity is to calculate variance inflation factors (VIFs) for each explanatory variable. This measures the degree to which the variance of a variable has been inflated because that variable is not orthogonal to other variables. If the VIF value for a variable is greater than 10, this suggests that the variable is close to being a linear combination of other explanatory variables. Five of the 17 variables in the 2012 model have VIFs greater than 10. They occur in three out of the five main effects for the route class and quality-of-service categorical variables, and in two out of the nine interaction effects involving the categorical variables. The key continuous variables—log capacity and log distance—do not appear to be collinearly dependent on other variables in the model.

Although strong multicollinearity can affect the interpretation of the coefficients and even entail reversal of expected sign, multicollinearity is not likely to be a major problem for forecasting if the pattern of multicollinearity in the explanatory variables does not change materially between the data used for estimation and forecast purposes. This assumption is implicit in assuming a similar structure across the data sets.

- *Misspecification*: Misspecification of a model might be due to adopting an inappropriate functional form (such as assuming a linear relationship between variables that are actually related nonlinearly) or due to omitted variables, for example. Misspecification can substantially affect coefficient estimates, for example, omitted variables will bias the

coefficient estimates. If nonlinearity is the source of the misspecification, then although the model will still be valid as linear approximation, it could not be expected to provide unbiased predictions. Misspecification is therefore an important problem. Two formal specification tests of the DAA model rejected the null hypothesis of no misspecification.⁸ Augmented partial residual plots suggested that the linearity assumption breaks down at the lower and upper ends of the range of values of outputs.

The misspecification issue is important because it implies that the model is unlikely to predict well for some types of services and the prices for those services may diverge from costs. The alternative specifications that were tested, which included adding higher-order terms for the capacity and distance variables, including additional conditioning variables and allowing for route-specific random effects, resulted in an improvement of the results of the specification tests. However, in only one of the two tests of misspecification was the null hypothesis of correct specification accepted. The RESET tests, which is particularly oriented to the problem of omitted variables was not satisfied.

The problem of omitted variables in relation to the 2014 dataset was raised by one of the stakeholders in comments on the draft report, and the same point applies to the 2011 dataset. Almost certainly there are relevant variables not available in the dataset. Potentially relevant variables of this kind include: (a) data relating to buyers such as buyer size and the number of contracts the provider has with the buyer; (b) whether the provider is using its own facilities or not. However, this issue cannot be feasibly addressed within the current review because we only have the data that is available. We are unable to assess the relative importance of this problem. In short, characteristics of the dataset limit the scope to remedy some of the issues raised by the diagnostic tests.

4.3 Application to 2014 data

The DAA 2012 model and the same two alternative specifications (additional variables and a random effects specification) were then estimated using the 2014 data. This provides an important insight into whether the model specification is structurally stable over time. The results are presented in Annex C.

We found that the DAA 2012 specification does not perform well when applied to the 2014 data. The goodness-of-fit is much lower: with an R^2 of 0.643 (compared to 0.842 when applied to the 2011 data) and a root-mean-squared-error (RMSE) of 0.56 (compared to 0.44 previously).

There are changes in the signs of some of the variables. The coefficient on the protection variable is inconsistent with the expectation that providing protection involves some additional cost. One interpretation might be that protection tends to be available on routes where it can be more easily provided, but the change from the 2011 data would be difficult to explain, and furthermore, the protection coefficient has a positive sign on both of the two alternative specifications shown in Annex C.

⁸ Link test for specification of dependent variable and RESET test for omitted variables, functional form and correlation between explanators and residual.

There are also changes in signs of the QOS indicator variables, which are more consistent with expectations, because their coefficients are negative and the absolute values of the coefficients on QOS 3 and 4 are greater than for QOS 2 (and QOS 1, which by implication is zero). However, these main effects cannot be accurately interpreted without taking into account the 9 interaction terms. There are some changes in the signs and magnitudes among the interaction terms between quality and route class and between quality and capacity.

Another notable result of estimating the DAA 2012 specification using the 2014 data relates to the magnitude of the coefficients on the capacity and distance variables. Together these two coefficients can be interpreted as an index of economies of scale, where a value of 1 represents constant returns to scale and a value less than 1 indicates economies of scale (where scale refers to the scale of a specific contract). In the DAA 2012 model these two terms added to 0.82, but when estimated with the 2014 data they are considerably smaller summing to 0.44 (although still highly statistically significant). This essentially means there has been a substantial change in the slope of the relationship between charges and the key output variables, capacity and distance.

The diagnostic statistics for the DAA and alternative models estimated with 2014 data are shown in Annex C. The same general observations relating to non-normality and heteroscedasticity of the residuals and the significant presence of outliers and observations with a high degree of influence apply to these models when estimated with the 2014 data. The following points are notable:

- Severe outliers are more frequent in the 2014 data. For the DAA 2012 specification, severe outlier residuals represent 0.75 per cent of the sample, compared to 0.58 per cent when estimated with 2011 data.
- There is a higher degree of multicollinearity between the regressors, with 8 out of the 18 coefficients in the DAA 2012 specification having VIF scores greater than 10 (compared to 5 out of 17 previously).⁹
- The two test statistics for misspecification shown in Table C.4, namely the RESET test and the link test, continue to strongly reject the null hypothesis that the model is correctly specified and have deteriorated for the DAA 2012 specification.

The two alternative specifications, when estimated with the 2014 data, represent a considerable improvement over the DAA 2012 specification. They have considerably better fit, whereas previously with the 2011 data, they provided only a marginal improvement in goodness-of-fit. For R^2 the 'Additional variables' model is 0.689, and for the 'Random effects' model is 0.678, which compare favourably to the R^2 of 0.643 for the DAA 2012 specification.

The higher-order output terms and the additional variables are statistically significant, and more strongly so than for the corresponding models using the 2011 data. The coefficients on the main effects on log capacity and log distance sum to 0.57 and 0.58 for the 'Additional variables' and 'Random effects' models respectively, which is greater than the sum of the effects on these variables in the DAA 2012 specification (0.44).

⁹ The additional variable relates to the categorization of the interface variable with 3 types rather than 2 types.

The route-class variables are positive in these alternative models, and the coefficient on the regional route type is greater than for the Metro route type. These implies an ordering of cost between Inter-capital, Metro and Regional route types from lower to higher, which is more meaningful than the coefficients in the DAA 2012 specification. However, this interpretation is again complicated by the interaction terms.

The diagnostic statistics in Table C.4 show that the observations relating to non-normally distributed and heteroscedastic residuals, and the relatively large number of influential data points, apply equally to the alternative models. There also continues to be a high degree of multicollinearity. But these two models do perform better in terms of the misspecification tests. Although the RESET test continues to reject the hypothesis of a correctly specified model, and suggests there are important omitted variables, the link test is satisfied for both of these models.

In summary, the DAA 2012 specification performed much more poorly with the 2014 data while the two alternative models improved considerably on the DAA 2012 specification when the 2014 dataset was used. These observations support a conclusion that further investigation of the most appropriate modelling specification to use with the 2014 is warranted. That is the subject of chapter 5.

4.4 Estimating with the pooled 2011 and 2014 data

The DAA 2012 specification was also estimated using the combined data for 2011 and 2014. An additional variable, ΔT , was added to the model, taking a value of 0 in 2011 and 3 in 2014. The results from this work were reported in the workshop paper (Economic Insights 2014). The results from this work indicated that the ΔT term was highly statistically significant and indicated that, all other factors held constant, annual charges declined by approximately 10 per cent per year on the deregulated routes (based on OLS estimates). Tests also confirmed that the hypothesis of no change in the coefficients collectively is rejected.

During discussions at the workshop of 24 April 2015 it was decided that, given the rapid changes in the market and the incompleteness of the 2011 dataset, it would be preferable to develop the econometric model using only the 2014 dataset, rather than a pooled data for 2011 and 2014.

5 DEVELOPING A PREFERRED MODEL

This section addresses the requirement to develop a recommended model to calculate DTCS prices on declared routes based on regression equations which best fit observed prices on deregulated routes. The methodological approach and the analytical steps to develop a preferred model are explained.

5.1 Methodology

The methodology for determining an appropriate statistical model needs to have regard to the objective. The objective, in this case, is to develop a model based on deregulated routes that have been determined to be sufficiently competitive that they can be used to establish benchmark prices to help set regulated prices for declared routes.

Given the objective of determining benchmark prices for regulatory purposes, the model needs to be both a reasonable predictor of prices for deregulated routes but also transparent, simple and relatively easy to apply in a regulatory context. If the model is not a reasonable statistical representation of the price generation process there will be a low level of statistical confidence that it can produce benchmark prices that are appropriate for regulatory purposes. As the model will be of interest to a wide range of stakeholders and will need to be used regularly in various contexts it will need to be transparent and reasonably parsimonious in terms of the inputs that are required.

However, in order to meet both objectives it is necessary to undertake considerable econometric testing to identify the most important explanators and then simplify to achieve a transparent and practical model to determine relevant price benchmarks.

The preferred strategy for model building, for the first stage of the methodology, is to start with a general specification and move to more simple specifications, in part to minimise the risk of omitted variable bias, and also to make effective use of hypothesis testing in the specification search (Greene 2012 p.178). The specification search can then use hypothesis testing and goodness-of-fit comparisons to narrow down the model to one which includes the most important variables.

Hypothesis testing is concerned with addressing questions such as: whether an individual variable can be considered as relevant, in the sense that it has a significant influence over the dependent variable, and therefore should be included in the model; whether one functional form provides a better representation of the underlying data generation process than another; whether the estimated values of certain parameters are consistent with a theoretical model; and/ whether the regression model is adequate overall. Testing hypotheses involves ascertaining the degree of confidence with which certain restrictions can be imposed on the values of parameters of the model. For example, we would want to test the null-hypothesis that parameters of quadratic and interaction terms are equal to zero, and ensure that any other proposed restrictions imposed on a more general specification are supported by hypothesis tests. In situations where competing models cannot be represented as special cases of a more general specification, it may not be feasible to rely on hypothesis tests for the purpose of model selection, and reliance is usually placed on goodness-of-fit comparisons.

Once a preferred model has been identified based on statistical testing and supplemented by judgements about relevant economic considerations a further stage of analysis can introduce simplifications to develop a final model for determining relevant price benchmarks for regulatory purposes. The price benchmarks are likely to be best interpreted as a reference point for regulatory purposes and may need to be supplemented by other information to set specific regulatory prices.

5.1.1 Economic & Econometric Specification

It is useful to firstly clarify the *economic* relationship the econometric model is intended to represent. The ACCC indicated that the prices on deregulated routes should broadly reflect costs and include a normal rate of return on investment (ACCC 2012 p.7), and accordingly we interpret the proposed regression function as a cost function.

In economic theory, the cost function is the minimum cost of supplying different levels of outputs given the prevailing input prices. In a cross-sectional study, movements in input prices are not present, and if we assume that different providers face the same factor prices then input prices are constant within the sample, and the explanatory variables of the cost function are the outputs and other relevant explanatory variables that affect costs (hereafter referred to as conditioning variables). If the data extends over time, then movements in input prices may also be relevant explanatory variables and there is also the possibility of technical change.

We treat capacity and distance as the outputs associated with the contract, and all other variables are treated as conditioning variables, whether they are features of the contract (such as term, or start date, or route type) or factors external to the contract (such as the provider's total throughput under all contracts on the same route, or the total throughput of all providers on a route). This distinction aids conceptually, but is not expected to be especially important in the econometric analysis.

The log-linear specification used in the 2012 study, when viewed as a cost function, would correspond to a Cobb-Douglas cost function. This functional form may be unnecessarily restrictive. DAA observed that “predictions at the lower range of values are relatively accurate” but “at the upper tail of the annual charges perform relatively poorly”. Our review of the DAA model has confirmed that there is some nonlinearity in the relationship between the log of costs and the log of outputs. The most commonly used flexible form of cost function that includes quadratic and interaction terms is the translog specification, although there are other flexible functional forms available (Chambers et al 2013). However, the translog function is widely used and understood, effective in a wide range of situations, and relevant in the context of the overall objective of developing a transparent and relatively simple model.

Most studies use a second order translog function, but in line with the general-to-specific methodology and to ensure that the functional form is capable of capturing the nonlinearity in the relationship between outputs and cost, this analysis commences with a third-order translog cost function specification. The third-order translog cost function with a single input and multiple outputs can be written as (Said 1992):

$$\begin{aligned}
 (5.1) \quad \ln C(w, \mathbf{y}) &= \gamma_0 + \gamma_w \ln w + \sum_i \gamma_i \ln y_i + \sum_i \gamma_{wi} \ln w \ln y_i \\
 &+ \frac{1}{2} \gamma_{ww} (\ln w)^2 + \frac{1}{2} \sum_i \gamma_{wwi} (\ln w)^2 \ln y_i + \frac{1}{2} \sum_i \sum_j \gamma_{ij} \ln y_i \ln y_j \\
 &+ \frac{1}{2} \sum_i \sum_j \gamma_{wij} \ln w \ln y_i \ln y_j + \frac{1}{6} \sum_i \sum_j \sum_h \gamma_{ijh} \ln y_i \ln y_j \ln y_h \\
 &+ \frac{1}{6} \gamma_{www} (\ln w)^3
 \end{aligned}$$

Here, C is total cost, w is an index of input prices, y_i represents the quantity of output i and the Greek symbols denote unknown parameters to be estimated. Given the assumed non-variation of input prices between providers, the base-year of the index w can be chosen to be the current year, so that $\ln w = 0$. This can be used to simplify equation (5.1), but we also extend it by introducing conditioning variables (z_k). To limit the proliferation of variables, the z 's are included in the specification only with main effects and interactions with the output variables.

$$\begin{aligned}
 (5.2) \quad \ln C(w, \mathbf{y}) &= \gamma_0 + \sum_i \gamma_i \ln y_i + \frac{1}{2} \sum_i \sum_j \gamma_{ij} \ln y_i \ln y_j \\
 &+ \frac{1}{6} \sum_i \sum_j \sum_h \gamma_{ijh} \ln y_i \ln y_j \ln y_h + \sum_k \beta_k z_k + \frac{1}{2} \sum_i \sum_k \beta_{ik} \ln y_i z_k
 \end{aligned}$$

In the present context, we adopt the assumption that there are two outputs, capacity and distance, and the remaining relevant variables are treated as conditioning variables (z 's). Therefore, equation (5.2) can be written more specifically as shown in equation (5.3). This represents the general approach to functional form adopted as the starting point of the econometric analysis.

$$\begin{aligned}
 (5.3) \quad \ln C(w, \mathbf{y}) &= \gamma_0 + \gamma_1 \ln y_1 + \gamma_2 \ln y_2 + \frac{1}{2} \gamma_{11} (\ln y_1)^2 + \frac{1}{2} \gamma_{22} (\ln y_2)^2 \\
 &+ \gamma_{12} \ln y_1 \ln y_2 + \frac{1}{6} \gamma_{111} (\ln y_1)^3 + \frac{1}{2} \gamma_{112} (\ln y_1)^2 \ln y_2 + \frac{1}{6} \gamma_{222} (\ln y_2)^3 \\
 &+ \frac{1}{2} \gamma_{221} (\ln y_2)^2 \ln y_1 + \sum_k \beta_k z_k + \frac{1}{2} \sum_k \beta_{1k} \ln y_1 z_k + \frac{1}{2} \sum_k \beta_{2k} \ln y_2 z_k
 \end{aligned}$$

The key strategies for the specification search in relation to the functional form are:

- to test, for each of the conditioning variables, whether the interaction terms with output should be excluded from the model, and if so, whether the main effect should be retained or excluded; and
- to test whether the output terms can be simplified, such as whether the third-order expressions should be excluded, or whether a single index of output can be used.

These decisions are aided by reference to:

- hypothesis tests of the individual or joint significance of coefficients or sets of variables;
- algorithms for excluding variables within a general-to-specific modelling method; and
- other tests including goodness-of-fit comparisons, and consistency with the economic theory underlying the cost function.

The specification search is explained in section 5.1.3 and more fully documented in Annex D. This process led to a simplification of the cost function to a second-order translog form in which the conditioning variables enter linearly, without any interaction terms with the outputs. This specification can be expressed as:

$$(5.4) \quad \ln C(w, \mathbf{y}) = \gamma_0 + \gamma_1 \ln y_1 + \gamma_2 \ln y_2 + \frac{1}{2} \gamma_{11} (\ln y_1)^2 + \frac{1}{2} \gamma_{22} (\ln y_2)^2 + \gamma_{12} \ln y_1 \ln y_2 + \sum_k \beta_k z_k$$

5.1.2 Variables and Expected Signs

The initial steps of the analysis involve estimating the general model described in equation (5.3). Annex A discusses all of the variables available in the 2014 dataset. Some of the variables were not included in the analysis because either:

- Advice provided by stakeholders at the workshops indicated that they were not related to transmission services or transmission pricing, or
- There was high correlation between variables that were essentially slightly different ways of measuring the same thing.

This section summarises the conditioning variables available in the dataset that were included in the most general models discussed in Annex D, and discusses the expected signs. These variables were all initially included in the general models including the interaction terms with outputs implied by equation (5.3).

In several cases economic theory does not provide a clear indication of what the sign of the estimated regression coefficient should be. The capacity and distance variables are expected to have positive signs but higher order terms could be negative or positive, indicating a diminishing effect or an intensification effect respectively, as capacity and distance increase. These effects may arise if scale effects are not fully exploited when there is competition, or where the model has effectively captured the competitive process. An econometric model can usually be only expected to be valid locally, and it is not valid to extrapolate the model to limiting cases far outside the domain of the sample.

The conditioning variables and their expected signs are:

- *Log # suppliers*: Number of firms with their own transmission infrastructure within 150 metres of a Telstra exchange at the A-end and B-end ESAs summed and divided by 2. A negative sign is expected if more suppliers means more competitive pressure to lower price but a positive sign could arise if economies cannot be realised.

- *Log # DTCS providers*: Number of providers of DTCS services on a given route. This differs from the foregoing measure, which includes providers who own the transmission facilities at relevant ESAs but are not providing DTCS services to third parties on the route in question. A negative sign is expected if more providers means more competitive pressure to lower price but a positive sign could arise if economies cannot be realised.
- *Log route throughput*: Aggregate capacity (in Mbps) of all contracts supplied on a given route by all providers represented on that route. Route throughput could be related to economies of scale to the extent that DTCS providers supply services using shared facilities. This would imply an expected negative sign.
- *Log provider-route throughput*: Aggregate capacity (in Mbps) of all contracts supplied on a given route by a single provider. Provider route throughput could measure route economies of scale for the relevant provider to the extent that the provider supplies all of its services on the relevant route using its own facilities. If so a negative sign is expected.
- *Log ESA throughput*: The sum of the reported capacity of every contract on routes with the relevant A-end or B-end ESA. The ESA throughput measure may indicate demand pressure or capacity constraints, if exchanges within ESAs that have higher traffic density are also those that require more frequent capacity augmentation. To the extent that high ESA throughput reflects either demand pressure or capacity constraints it would be expected to have a positive effect on DTCS prices.
- *Log # DTCS services*: Total number of contracts supplied on a given route by all providers of DTCS services on that route. This differs from route throughput as it does not take into account the capacity of the contracts, only the number of such contracts. Like route throughput, it could be related to economies of scale to the extent that DTCS providers supply services using shared facilities. It may also be related to competitive pressure. In both cases this would imply an expected negative sign.
- *Contract start date*:¹⁰ The date at which the relevant contract commenced. This data is believed to have some inconsistencies relating to whether the contract renewals have been recorded. A negative sign is expected given that advances in technology and competition are expected to lead to price reductions over time, although the ability of providers to vary prices through the term of a contract may vary. At the workshops in April 2015 technical experts expressed concerns about how to measure the contract start date and also the accuracy of the data. If contract prices do not vary within their term, then the coefficient on contract start date would be approximately equal to the average percentage daily rate of change in prices. However, the data quality and price variation issues mentioned mean that this interpretation is open to question.
- *Contract term*: . It is not clear how contract term affects risk and the pricing of contracts and views of stakeholders differed.
- *Protection*: This refers to the existence of a back-up service in the event of an interruption. There are two types of protection indicated in the dataset, geographic

¹⁰ Contract start date is expressed as a Stata date, which is defined by the number of days from 1/1/1960 to the date in question.

diversity and electronic protection. Very few were of the latter type.¹¹ The existence of protection is expected to have a positive sign.

- *Route class*: Indicates whether the route is inter-capital, metropolitan (i.e. between ESAs in the same capital city), regional (including between metropolitan ESAs and regional ESAs) or tail-end (where both the A-end and B-end are in the same ESA). It is not clear what the signs should be for the metro variable but a positive sign is expected for the regional route class reflecting higher costs.
- *The ACCC's Quality of Service (QOS) classification*: To the extent that the ACCC quality of service measures reflect quality, the coefficients should have negative signs relative to the default measure of highest quality with the coefficient increasing in absolute size with lower quality.
- *Interface type*: The type of interface technology, which may be Ethernet, SDH, EoSDH. Different interface may have different quality of service. Newer interface types may have a lower cost of supply. Ethernet is the increasingly preferred interface type on unregulated routes.

At the workshops participants raised concerns about the extent to which the data and proposed model adequately capture the competitive process, particularly where there is considerable bundling of service offers. The ACCC has confirmed that the prices in the data set relate to prices that were actually paid by the customer and that it is difficult to remove all offers associated with a bundled negotiation. It is impractical to address bundling any further at this stage. Some of the other possibly relevant factors, for which there is no data, include: (a) data relating to buyers such as buyer size and the number of contracts the provider has with the buyer; and (b) whether the provider is using its own facilities or not. Again, it is not possible to test for the effects of variables for which no data available.

5.1.3 Estimation method

Outliers

A feature of the data sample is the presence of severe outliers and the possibility (or likelihood) of resulting bias in coefficient estimates obtained using ordinary least squares (OLS) regression. At the initial stages of the analysis in which the DAA 2012 model and the variations on that model were estimated using the 2014 data, as discussed in chapter 4, the most severe outliers and observations with high leverage and influence were identified and lists of those observations were provided to the ACCC. The ACCC requested the information providers to check these records and in several cases the data was amended by the information provider, and in a few cases, where an observation was found to be in error but could not be rectified, the information provider indicated that the observation should be dropped from the dataset.

¹¹ On unregulated routes, out of 6,767 services, 3,883 were geographically protected and 83 were electronically protected.

Given the size of the sample and concerns expressed at the workshop about removing outliers, rather than making judgements about specific individual observations, the following estimation methods that limit the influence of outliers were used:

- quantile regression at the median, which is based on least absolute deviation (rather than minimum squared deviation) and is therefore less affected by extreme values; and
- robust regression, which refers here to an algorithm available in Stata (*rreg*) for iterative least squares estimation in which high valued residuals are down-weighted. The robust regression method available in Stata is only one of dozens of different methods of this kind, which may make this approach less attractive than the quantile regression approach.

Unobserved factors

A second issue relates to limitations of the set of variables available to measure attributes that are relevant to pricing, and the possibility (or likelihood) that there are important unobserved or latent variables, particularly route-specific factors that affect the cost function. The facilities on different routes may be of different ages, use slightly different technologies, or achieve different performance levels given the environment in which they are situated or the nature of the traffic on them. To address the issue of unobserved variables, we also test, as an extension of the basic OLS model, two specifications that seek to isolate the route-specific factors, namely the fixed effects and random effects methods.¹²

The strategy used for choosing between these models is as follows. Statistical tests include the F-test, used to test the significance of fixed effects versus no fixed effect (OLS), and the Hausman test, used to choose whether the random effects model is valid relative to the fixed effects model. There are also important issues relating to how the model is to be used when applied out-of-sample to estimate competitive cost benchmarks on regulated routes.

The review of the 2012 model also highlighted the likelihood of heteroscedastic residuals. Although there are methods for robustly estimating standard errors in this context, this can make specification tests, such as the Hausman test, more difficult. In addition, given that a relatively high statistical threshold is used for selecting relevant variables there is less concern about the impact of potential heteroscedasticity on statistical tests and in any case the presence of heteroscedasticity does not by itself signal bias in the estimated coefficients. For this reason, corrections of this kind are used sparingly when developing a preferred model. This approach was supported by experts at the technical workshop on 24 April 2015.

Out of sample testing

A random sub-sample of 677 observations, or 10 per cent of the full sample, was omitted from the data used for model estimation, and instead used for model validation during the

¹² Fixed and random effects estimation refers to particular methods of panel data analysis which seek to identify a time-constant unobserved effect. Random effects estimation treats the unobserved effect as part of the stochastic term of the model, a random variable that takes a different value for each panel has a single value within each panel. Fixed effects estimation treats the unobserved effect as differences in the intercept for each panel.

specification search. For each of the models presented in this section, the main goodness-of-fit measures, namely the root mean squared error (RMSE) and the mean absolute error (MAE) are calculated separately for the estimation data and the validation sample.¹³ The same observations are used as the validation sample in each case.¹⁴

In the final analysis, once the preferred model was selected, it was re-estimated using all of the observations for contracts on unregulated routes — i.e. including the observations previously put aside to form the validation sample. This provides greater clarity and ease of replication, since otherwise the precise results would depend on the specific observations that were randomly chosen for the validation sample.

Centering data values

The data used in this analysis was centered, in the first place, before estimation. Each observation on each variable the data is transformed by subtracting the sample mean for that variable. For some analysis, as discussed below, the data is further transformed using the *xtdata* routine to facilitate specification search.

The approach to centering or mean correction of the data was raised at the technical workshop. It was suggested that nonlinear calculations involved in forming explanatory variables should be carried out before centering the data. We subsequently adopted this approach to centering favoured by the stakeholders. However, note that this is not standard practice in applied production econometrics involving the translog form where the aim of mean correction is to mean-correct the base data so as to change the units of measurement of this data so that the sample means become equal to 1. With this approach the first order coefficients may then be interpreted as elasticities at the sample means (because the log of 1 is 0) and hence the economic plausibility of the estimated coefficients may be readily assessed.

In the final analysis, once the preferred model was selected, it was re-estimated using uncentered data. This aids transparency and avoids the need to calculate the intercepts in forming the pricing model.

Goodness-of-fit measures

One way of comparing models is in terms of their goodness-of-fit both within sample and out-of-sample (using the validation sample). It is not a conclusive criterion because the alternative methods are motivated by a trade-off between reliability of coefficient estimates for prediction and goodness-of-fit within sample. One goodness-of-fit measure we report is the RMSE and another is the MAE. A third goodness-of-fit measure is the correlation between the fitted and actual values of the dependent variable, or R^2 . For some models we also report the Bayesian Information Criterion (BIC), although this is not available for all models.

¹³ Denoting the residual as e , the RMSE is defined as: $\sqrt{\left(\sum_{i=1}^n e_i^2\right)/n}$. The MAE is defined as: $\left(\sum_{i=1}^n |e_i|\right)/n$.

¹⁴ This is ensured by using the same seed for generating the random number used to select the validation sample.

Normality of residuals

Although normality of the residuals is not essential to obtain unbiased estimates of the regression coefficients, substantial departures from normality, including the presence of severe outliers, can affect coefficient estimates and standard errors, and reduce the effectiveness of a model needed for prediction. One test of the normality of the residuals is the Doornik-Hansen test of the null hypothesis that the residuals are distributed normally.¹⁵ Another test of normality is the frequency of severe outliers. The IQR (interquartile range) test identifies severe outliers as a percent of the sample.¹⁶ All of the models tested in this study fail the tests of normally distributed residuals, and the tests indicate that the nature of the non-normality is relatively heavy tails.

General to specific approach

A preferred strategy for econometric model building is to start with a general specification and move to more simple specifications, in part to minimise the risk of omitted variable bias, and also to make effective use of hypothesis testing in the specification search (Greene 2012 p.178).

In situations where competing models cannot be represented as special cases of a more general specification, it may not be feasible to rely on hypothesis tests for the purpose of model selection, and reliance is usually placed on goodness-of-fit comparisons, as discussed above.

In the general-to-specific approach an initial model is formulated that expresses the economic relationship being estimated in its most general form and encompasses all of the variables and effects of interest. The general economic model that serves as the starting point for the general-to-specific analysis is the third-order translog cost function. Six models were initially tested to resolve the issue of the appropriate estimation methods. These are:

- (1) OLS
- (2) Quantile regression
- (3) Robust regression
- (4) Fixed effects model (with route-specific effects)
- (5) Random effects model (with route-specific effects)
- (6) Quantile regression (with data transformed for random effects estimation).

The results of the initial analysis are used to inform the exclusion of variables, and simplification of the functional form, in the second round. Tests of the individual and joint significance of parameters in the first round model are one important criterion for this assessment. In addition, a general-to-specific methodology was implemented using the user-written Stata program, *genspec*, which derives a more parsimonious model by removing

¹⁵ Implemented with the user-written routine: *omninorm*.

¹⁶ Implemented with the user-written routine: *iqr*. The IQR = 75th percentile – 25th percentile. A severe outlier is either less than: 25th percentile – 3 × IQR; or is greater than: 75th percentile + 3 × IQR. Severe outliers represent only about 0.0002 per cent of a normal distribution.

variables which have least influence on the Bayesian Information Criterion (BIC), among other tests (see: Clarke 2014). This procedure is only available for the estimation methods based on least squares and is not available for quantile regression. The general-to-specific (GETS) algorithm provides another perspective on the variables that are candidates for excluding from the model. We regard these two approaches (namely the joint significance tests of parameters and the GETS procedure) as complementary, rather than alternatives, and use both methods at different stages of the analysis.

Parsimony

As emphasised by both the ACCC and stakeholders, there is a positive benefit to simplicity or parsimony in the model, and this requires a trade-off between simplicity and goodness-of-fit. This has been approached in two ways. During the specification search using the general-to-specific methodology a relatively high standard of statistical significance has been used. Usually t-statistics of 3 are required as a minimum. Secondly, in the final stage of simplifying the model, some effects with higher levels of significance have been removed, such as interaction effects on conditioning variables, to achieve greater parsimony. This is intended not only to make the resulting pricing model easier to implement and use by regulated businesses, but also a suitable degree of simplicity can improve the predictive performance of econometric models (see contributions in: Zellner et al. 2004).

5.2 Deriving the preferred model

The process of deriving the preferred model is described in Annex D. This process of deriving this model is based on conventional hypothesis tests relating to model specification and the significance of coefficient estimates, whether individually or as part of groups. It also relied on a general-to-specific modelling procedure to aid in the simplification of the model by removing unimportant variables. The results of the parameter tests and the general-to-specific models are discussed in Annex D. See also the draft report (Economic Insights 2015a).

At the draft report stage, the quantile (median) model and the random effects model were preferred over the OLS model. The quantile model better deals with the problem of severe outliers in the data, whereas the random effects model identifies the effects of unobserved time-invariant route-specific effects in the costs structure. Major comments on the models presented in the draft report and the changes that have been made in response to those submissions are briefly summarised in the sections below.

5.2.1 Submissions to Draft Report

This section provides a brief summary of some of the major issues raised by industry stakeholders and technical experts, especially those raised in response to Economic Insights' draft report. More specific issues raised by these stakeholders are discussed in Annex D.

Some stakeholders were critical of the approach to model specification used by Economic Insights in the workshop paper, especially the undue use of economic intuition. Advice from industry specialists would assist to understand the potential economic relationships between variables. It was recommended to employ a more systematic econometric specification method, starting with a relatively short list of variables, some of which will be essential and

some optional, and using a well-defined fit criterion and other rules to narrow the model down to one that is simple. It was also observed that a model so derived will be unlikely to satisfy diagnostic tests because, given its simplicity, it will only capture broad features of the data, with the aim of providing a benchmarking formula that is useful for prediction and readily understood. In submissions to the draft report, some stakeholders indicated that the approach taken by Economic Insights had responded to these concerns and that the emphasis on moving to a final model should be on considering all of the elements of the estimation model in terms of how well it translates into an effective pricing formula.

Some stakeholders were concerned about the intrinsic volatility of prices in the dataset, which cannot be explained by regression analysis using the information available in the dataset. Some of the data issues included omitted variables (which we have noted), the implications of widespread bundling on the meaningfulness of the data, and the frequency of pricing at common price points. However, we note that the four most common price points on unregulated routes represent only 8 per cent of all contracts.

In regard to the concerns about data limitations, while we recognise that there are limitations, there is also a large number of observations and it is reasonable to assume that extreme effects will average out. If so, the econometric model may do a reasonable job of predicting prices. We do not have information that would identify the sign and size of potential bias in the price predictions, nor are we convinced that the predictions of the model will be materially biased or result in too high a price for DTCS services on regulated routes.

In regard to bundling, the ACCC has confirmed that the prices in the data set relate to prices that were actually paid by the customer and it would be difficult to remove all offers associated with a bundled negotiation. It is therefore impractical to address the issues raised about bundling in this review.

It was considered by some that there remains some market power in unregulated routes, which would distort efficient pricing. The model development process set out in Annex D shows that variables expected to be related to the degree of market power, such as the number of service providers on a route, were not found to be significant. One expert proposed that stochastic frontier analysis (SFA) be used, which is discussed in Annex D. A problem with the SFA approach in this context is that it would forecast lower prices based on an efficiency interpretation of the unexplained variation in the data, but given the scope of this variation, a premium would then need to be added to ensure prices were sufficient to finance investment and allow for estimation uncertainty. But it is not clear what the premium should be or how to calculate it. To allow for the possibility of some residual non-competitive effects in the deregulation data we tested for provider-specific fixed effects as an alternative and found they were significant but not suggestive of market power (since the largest providers were close to the centre of the distribution of effects) and could reflect a range of factors. The provider-specific fixed effects were adopted in the final model.

Several stakeholders felt that the model should be much simpler than those presented in the draft report. On the other hand, at least one of the experts suggested more complex econometric methodologies that are on the cutting edge of complexity, and in some cases developed for the present purpose — combining the General Additive Models (GAM) specification with the Robust random effects (RRE) and the Lasso method. As a general rule,

it is desirable that economic or econometric analysis for regulatory purposes should use well-established methods, and it is not clear that the proposed method has been used before in the econometric analysis of cost functions or other related contexts. They are more experimental in nature, and in our view, not sufficiently well established in this type of analysis to warrant their adoption.

There were also concerns related to tail end services. Firstly, it was observed that contracts for 2 Mbps services were largely associated with connectivity and a high proportion of these contracts included a bundled tail end service. As discussed in Annex D, we have attempted to address the question as to whether there is a ‘bundled tail end’ effect by introducing into the model some effects relating to services with 2 Mbps capacity, firstly by using an indicator variable for contracts with 2 Mbps capacity, and secondly using the same indicator variable together with an interaction term between this indicator variable and the variable ‘log distance’. In the second of these two approaches, the main effect and the interaction effect are highly correlated with each other but yield coefficient estimates of opposite sign. When only the main effect was included in the model, its coefficient was not significantly different from zero. Therefore, we could not identify a bundled tail-end effect. The second issue relating to tail ends was disagreement with the method of setting prices for regulated stand-alone tail-end services using the econometric pricing model and assuming a 2 km distance. The issues related to tail end pricing are discussed in section 6.3.

There were also views and suggestions on which of the econometric models at that stage provided the best platform for further development, and a number of suggestions were provided on further simplifying the model, and specific methods to improve the accuracy or transparency of the analysis. These were largely adopted and are noted in Annex D.

5.2.2 Addressing issues raised in the submissions

The final development of the preferred model, presented in section 5.3, has benefitted from the stakeholder comments discussed in the foregoing section. We have also sought to correct any errors or misinterpretations identified by the submissions that we agree with. This section summarises how issues raised in submissions have been addressed.

The random effects model is now preferred to the quantile regression model because the quantile model was considered to be potentially unstable since it required a large number of iterations to converge. There are two options for estimating the random effects model in Stata, and the results obtained are quite similar. We use the maximum likelihood estimator of the random effects (MLE-RE) method because the standard error estimates for the stochastic components are considered to be more reliable, which is important for the adjustment from logs to levels when deriving the pricing model. Hereafter the MLE-RE estimator is used for all the models shown unless otherwise indicated.¹⁷

As discussed in Annex D, the final preferred models are estimated using uncentered data (i.e. as logs of their original units) to avoid the need to calculate the intercept and to improve transparency. They are also estimated using the entire sample of exempt routes contracts (i.e. including the part of the sample that was put aside for out-of-sample validation during the

¹⁷ The models have also been estimated using the validation sample and those results are available if needed.

specification search).

Various suggestions for simplification of the model have been adopted. These include removing the interaction terms on interface-type, contract term, contract start date, log ESA throughput and log route throughput. The contract terms was also dropped as it became insufficiently significant after these changes. We have also tested the exclusion of the contract start date variable due to concerns about the data quality. This has resulted in a much more parsimonious preferred model.

As mentioned, the ‘Tier 2’ variable as a measure of quality and the associated interaction terms have been replaced by provider-specific fixed effects. In part this addresses the argument put forward by one expert that market power may not be route-specific, but more broadly-based due to bundling practices, and we have sought to capture any effects of this kind through provider-specific effects. The provider fixed-effects would also capture differences in quality of service between providers, to the extent they are broadly based, and may also reflect differences in efficiency. Lastly, since small providers tend to have a relatively higher rate of extreme outliers, the provider-specific effects relevant to those providers may in part compensate for the effects of some of those outliers.

One of the stakeholders had concerns about the interpretation of the coefficients on *ESA throughput* and *route throughput*. We suggested in the draft report that the negative coefficient on route throughput may reflect economies of scale in DTCS infrastructure if providers share facilities. In regard to the interpretation of the positive coefficient on *ESA throughput*, we suggested in the draft report that this may be due to capacity constraints at exchanges in ESAs with higher density telecommunications traffic, particularly if the traffic at those ESAs also has relatively higher growth rates. However, we recognise the need to be cautious when making interpretations of this kind, particularly since it is likely that wholesale transmission throughput is small relative to the amount of self-supplied transmission traffic, as the stakeholder pointed out. Given these issues of interpretation we tested the model with these two variables excluded and found that, although they were individually and jointly statistically significant. Their removal had little effect in reducing the goodness-of-fit of the model.

A number of suggestions made by the industry stakeholders and experts in relation to transforming the econometric model into a pricing model. These include issues such as: the appropriate method of calculating the Jensen’s inequality adjustment; and the need to present prediction confidence intervals for the pricing model and report predicted prices for tail end services. These issues are discussed in chapter 6.

5.3 Preferred models

This section presents the preferred econometric model to characterise the determination of the prices of DTCS services on unregulated routes. Chapter 6 addresses the application of this model for the purpose of making predictions of benchmark competitive prices on regulated routes.

Table 5.1 shows three models. The first of these two models includes the contract start date variable, and the second excludes that variable. In the first model the coefficient on the

contract start date is equal to -0.00005 , and this coefficient can be loosely interpreted as a daily rate of change in prices in percentage terms, which is equivalent to approximately -1.8 per cent per year. However, issues were raised about the quality of the data for contract start date, which casts doubt on the reliability of this estimate of the annual rate of price decline.¹⁸ The second model shows the effect of excluding this variable. Because of uncertainty about the quality of the data for contract start dates, there is a preference to exclude this variable.

The second model shown in Table 5.2 is a 2nd order translog model in two outputs, capacity and distance, and with five conditioning variables, namely:

- route throughput
- ESA throughput
- a route class effect
- a provider specific effect, and
- an interface type effect.

Using similar notation to that used in equation (4.1) the preferred model can be expressed by the following equation:

$$(5.5) \quad R = \alpha_0 + \alpha_1 C + \alpha_2 D + \alpha_{11} C^2 + \alpha_{12} C \cdot D + \alpha_{22} D^2 \\ + \alpha_3 \text{interface} + \alpha_4 \text{ESA t'put} + \alpha_5 \text{route t'put} \\ + \beta \cdot i.\text{route} + \varphi \cdot i.\text{provider}$$

where: R is the log monthly charge in \$, C is log capacity, D is log distance, *interface* is an indicator variable indicating whether the interface is SDH, *i.route* is a set of indicator variables representing the route categories, and *i.provider* is a set of provider fixed effects.

This model differs from the DAA 2012 model in the following ways. It includes higher-order terms on the outputs. Interface type is included in the model and protection is not included, and two additional conditioning variables are included. It does not include interaction terms between the conditioning variables and the outputs or interactions with other conditioning variables. Provider-specific fixed effects are used instead of the ACCC's QOS variable.¹⁹

The third model shown in Table 5.1 excludes the route throughput and ESA throughput variables. It can be expressed by the following equation:

$$(5.6) \quad R = \alpha_0 + \alpha_1 C + \alpha_2 D + \alpha_{11} C^2 + \alpha_{12} C \cdot D + \alpha_{22} D^2 \\ + \alpha_3 \text{interface} + \beta \cdot i.\text{route} + \varphi \cdot i.\text{provider}$$

¹⁸ It is considered to be more likely to understate the annual rate of price decline than to overstate that rate.

¹⁹ One of the models DAA included an alternative model with provider-specific fixed effects rather than the QOS variable, but the specification that included the QOS variable was preferred by the ACCC for applying to regulated DTCS pricing.

**Table 5.1: Random effects models
(2015 data, full sample, ML-RE estimator)**

<i>Predictor</i>	Model (1) incl. contract start date		Model (2) excl. contract start date		Model (3) excl. contract start date, route t'put & ESA t'put	
	<i>Coeff.</i>	<i>t-stat.</i>	<i>Coeff.</i>	<i>t-stat.</i>	<i>Coeff.</i>	<i>t-stat.</i>
constant	5.68261	35.07	4.78394	44.67	4.95344	57.57
log capacity	0.49789	43.07	0.49225	42.52	0.49147	42.51
log distance	0.09831	4.14	0.09500	3.98	0.11703	4.97
0.5(log capacity) ²	-0.03596	-12.48	-0.03523	-12.19	-0.03503	-12.13
0.5(log distance) ²	0.01319	2.03	0.01369	2.09	0.01295	1.97
(log capacity)(log distance)	-0.00375	-2.47	-0.00366	-2.41	-0.00472	-3.17
log route t'put	-0.01791	-3.27	-0.01966	-3.57	.	.
log ESA t'put	0.03327	4.11	0.03027	3.72	.	.
contract start date	-0.00005	-7.36
route class 2 (Metro)	0.17166	2.33	0.17391	2.35	0.22019	2.98
route class 3 (Regional)	0.31876	5.48	0.31498	5.38	0.32620	5.54
Provider #1	[REDACTED]					
Provider #3						
Provider #4						
Provider #5						
Provider #6						
Provider #7						
Provider #8						
Provider #9						
interface-type 3 (SDH)						
$\sigma(u)$	0.3166		0.3187		0.32467	
$\sigma(e)$	0.4276		0.4291		0.42864	
Goodness-of-fit						
R ² *	0.6837		0.6819		0.6767	
BIC	9310.6		9355.7		9355.6	
RMSE (based on ue)	0.5253		0.5267		0.5316	
MAE (based on ue)	0.3707		0.3711		0.3767	
Joint significance tests						
	<i>chisq</i>	<i>p-value</i>	<i>chisq</i>	<i>p-value</i>		
2 nd order output terms (df = 3)	198.5	0.0000	189.8	0.0000	196.9	0.0000
Route classes (df = 2)	41.9	0.0000	39.9	0.0000	36.9	0.0000
Provider fixed effects (df = 8)	1072.1	0.0000	1086.7	0.0000	1087.5	0.0000

Source: Economic Insights estimation results.

Notes: * Squared correlation between fitted and actual dependent.

Economies of scale

The coefficients on the main effects and the higher-order effects of the outputs (log capacity and log distance), when taken together, imply there are cost economies associated with the scale of individual contracts. As previously mentioned, the route throughput variable may reflect additional cost economies of scale associated with the multiplicity of contracts on

routes, beyond the economies associated with providing additional outputs within the scope of a single contract.

The sum of the partial derivatives of log cost to each of the log outputs in the translog model is a measure of economies of scale (when less than 1) or diseconomies (if greater than 1). Each of the partial derivatives of log monthly charge (i.e. “cost”) to one of the log outputs is equivalent to the elasticity of cost with respect to that output. The overall measure of economies of scale measure is equal to the sum of these cost elasticities with respect to the outputs.

When the log outputs are measured relative to their means, each elasticity, when evaluated at the sample means, is equal to the coefficient of the translog cost function on the main effect of the relevant output. However, when the output variables are not measured in this way, it is necessary to calculate these partial derivatives (i.e. elasticities). Using the notation from equation (5.4) these partial derivatives are:

$$(5.6) \quad \frac{\partial \ln C}{\partial \ln y_2} = \gamma_1 + \gamma_{11} \ln y_1 + \gamma_{12} \ln y_2$$

$$(5.7) \quad \frac{\partial \ln C}{\partial \ln y_1} = \gamma_2 + \gamma_{12} \ln y_1 + \gamma_{22} \ln y_2$$

On exempt routes the sample mean value for $\ln y_1$ is 2.564238, and for $\ln y_2$ is 2.532687. The elasticities, calculated using equations (5.6) and (5.7), and using the sample mean values for $\ln y_1$ and $\ln y_2$ are shown in Table 5.2 for the models shown in Table 5.1.

Table 5.2: Economies of scale per contract

	Model incl. contract start date	Model excl. contract start date	Model excl. contract start date, route t'put & ESA t'put
Elasticity of C to y_1	0.3962	0.3926	0.3897
Elasticity of C to y_2	0.1221	0.1203	0.1377
Economies of scale index	0.5183	0.5129	0.5274

Source: Economic Insights estimation results.

The results in Table 5.2 show that cost economies in the scale of contracts is indicated in the preferred and alternative models. The economies of scale index, which is the aggregate elasticity of cost with respect to outputs, is close to 0.5, which indicates that a 1 per cent increase in all outputs results in an approximate 0.5 per cent increase in costs.

These results indicate a greater degree of economies of scale than implied by the DAA 2012 model. In that model there were no higher-order effects for outputs and therefore the coefficients applying to log capacity and log distance are equal to the elasticities of cost for these two outputs. Those coefficients were 0.62262 and 0.19864 respectively, and economies of scale are indicated since the sum of these two elasticities, which is 0.82126. Although the DAA model also implied economies of scale within contracts, the degree of economies was not as pronounced as the results of this study indicate.

Route class effects

The route class effects seem to have a reasonable ordering of values. They imply that DTCS services are at their lowest cost on inter-capital routes, and for given values of the other variables in the model, they will on average be about 17 per cent higher cost on metropolitan routes and about 31 per cent higher cost on regional routes.

These relativities differ from those in the DAA 2012 model. Changes are to be expected since there have been some important changes in the dataset, such as the deregulation of additional routes between 2011 and 2014, and strong market growth which has added to the density of some routes. It would be incorrect to make a comparison between these models based only on the main effects on the route class variables, which imply that metro routes had lower prices than inter-capital routes and regional routes had only marginally higher prices. In the DAA 2012 model, route classes also interacted with the quality-of-service (QOS) variables, and the marginal effect of a route class needs to take into account the average values of these interactions. Using the notation of equation (4.1), the marginal effect of a route class i calculated at the sample means is equal to: $\beta_i + \gamma_{i2}\bar{Q}_2 + \gamma_{i3}\bar{Q}_3 + \gamma_{i4}\bar{Q}_4$, where β_i is the main effect on route class i ; γ_{ij} is the coefficient on the interaction of route class i with QOS level j , and \bar{Q}_j is the sample mean value of QOS level j .²⁰ Using these means and the values of the coefficients of the DAA model reported in Annex C, the marginal effects were: 0.0237 for the metro route class and 0.1250 for regional routes. The results of this study suggest broadly similar relativities between metro and regional prices, but higher relative prices on these routes compared to inter-capital routes.

Provider fixed effects

The provider fixed-effects are almost all significant. [REDACTED]
[REDACTED] The joint parameter tests for the provider-specific fixed effects strongly reject the null hypothesis that the coefficients on these variables are all equal to zero. [REDACTED]

[REDACTED] This does not appear to support the claim that market power effects are important within this sample, unless such effects are subsumed within other effects, as suggested by some stakeholders.

[REDACTED]. Since it was found that outliers were disproportionately represented among the smaller providers, these results tend to suggest that the most extreme of the provider-specific fixed effects are capturing part of the influence of outliers and thereby assisting to correct for differences in data quality. Aside from this type of influence, the provider-specific fixed effects may reflect

²⁰ [REDACTED]

differences in efficiency, product differentiation, market power, or possibly other factors.

Interface type

We understand that Ethernet interfaces are increasingly used in preference to SDH interfaces, and are considered more efficient and cost effective. The positive coefficient on the interface type is consistent with this observation.

Route and ESA throughput

The negative coefficient on route throughput may result from economies of scale in DTCS infrastructure if providers share facilities. However, interpretations of this kind need to be made cautiously since the level of wholesale transmission throughput may be small relative to the level of self-supplied transmission traffic.

In regard to ESA throughput, we suggested in the draft report that the positive coefficient on this variable may be due to capacity constraints if exchanges in ESAs with higher density telecommunications traffic, an interpretation that was challenged by one stakeholder. Again, this highlights the need for caution when interpreting some effects in the model. The positive coefficient on ESA throughput suggests that, when the values of all other variables in the model are given, the supply price is higher in ESAs of higher demand.

5.4 Concluding comments

Annex D provides some more information including diagnostic tests on the models shown in Table 5.1. There remain issues of non-normality of the residuals and the possibility that some relevant variables are not available in the dataset. However, as stakeholders have emphasised, the emphasis should be on deriving a benchmarking formula which is simple and readily understood, captures the broad features of the data, and is useful for prediction.

Chapter 6 explains how the preferred econometric model in Table 5.1, which characterises price formation on deregulated routes, can be formulated as a pricing model for regulated services.

6 APPLYING THE MODELS TO REGULATED ROUTES

This section of the report addresses the following matters:

- Explanation of the operation and use of the model and of any adjustments required to set efficient prices in regulated routes to allow for differences in the characteristics of declared areas from competitive routes.
- Predicted prices on selected regulated routes and estimates of the standard forecasting errors for the models when some of the variables are set to constants.
- Discussion of the application of the model to setting prices for tail-end services.
- Discussion of whether the model can be applied to calculating an allowance for change in prices over the 2015 FAD period to reflect expected productivity and cost movements.

A separate spreadsheet model has been developed to facilitate use of the model in setting prices for regulated routes. It has been used to present the price benchmarks in Section 6.4.

6.1 Current pricing model

The Final Access Determination (FAD) No. 1 of 2012 (ACCC 2012a) specified an annual use-related charge and a connection charge. The use charge was based on the regression model in Table C.1 and defined as follows (DAA 2012):

$$(6.1) \quad \begin{aligned} \text{Annual Charge} &= \exp\{7.682 + 0.623\ln(\text{Speed}) + 0.199\ln(\text{Distance}) + c \\ &+ t\} \cdot \exp(\hat{\sigma}^2/2) \end{aligned}$$

where:

$c = 0.078$ for protected service and 0.0 for unprotected service

$t = 0.000$ for intercapital routes, -0.081 for metro routes and 0.052 for regional routes

Equation 6.1 includes an adjustment term: $\exp(\hat{\sigma}^2/2) = 1.102$ (DAA 2012) which is needed to adjust for Jensen's inequality when expressing the model in levels rather than logs. The quality of service impacts, including interaction effects in the model presented in Table C.1 were excluded. This means that the intercept term included the quality effects and the regulated price cap implicitly assumed the highest quality service.

6.2 Pricing models

In the present study, our preferred models are shown as Model 2 and Model 3 in Table 5.1. An alternative, which includes the variable 'contract start date' is presented as Model 1 in the same table. The following discusses the transformation of Model 2 into a price model for unregulated services. Then in section 6.2.2 the application of Model 3 is discussed. Section 6.2.3 discusses the use of the contract start date in Model 1.

6.2.1 Application of Model 2

Using similar principles as those used in the 2012 FAD for moving from the estimated model to the pricing model, the predicted values of monthly charges using Model 2 can be expressed as follows:

$$(6.2) \quad M = \exp\left\{4.7839 + 0.4923 \ln C + 0.0950 \ln D - 0.0176(\ln C)^2 + 0.0068(\ln D)^2 - 0.0037(\ln C \ln D) + 0.2434I + \delta + \theta\right\} \cdot \exp\left(\hat{\sigma}^2/2\right)$$

where: M is the monthly charge; C is the capacity; D is the distance; $I = 1$ for an SDH interface and 0 otherwise; δ is a constant that varies by route-type (as discussed in section 6.2.2 below); and θ is a constant in place of the provider fixed effect. When the model is used for prediction, θ is a constant that applies equally, irrespective of the actual provider. This term is discussed below.

Provider effect

Most of the provider-specific fixed effects are relatively closely bunched together, with some smaller players as outliers (either positive or negative). These more extreme valued fixed effects may control to some extent for atypical characteristics associated with some of the smaller providers and data quality issues relating to some information providers. Options for the ACCC to consider for the *pricing model* would be to use the largest provider as the base (which by construction has a fixed effect equal to zero), since it represents the median or make a positive or negative adjustment, depending on the ACCC's view about factors not incorporated in the models.

In the following calculations we assume that the median value of the provider effect is chosen, in which case $\theta = 0$. We are not asserting that this value should necessarily be adopted.

Adjustment for route-type

The constant term that varies by route type, δ , is calculated in Table 6.1. It combines the effects of the route-class coefficients as well as the effects of route throughput and ESA throughput. These effects have been calculated using the mean values of route throughput and ESA throughput on regulated routes, to which the model will be applied.

There is a question as to how these route-specific effects should be calculated for stand-alone tail ends. It would not appear to be appropriate to assume that the route-class effect applying to tail ends should be zero, since that value applies to inter-capital routes. Instead it may be appropriate to apply the metro route class effect to ESAs located in capital cities and the regional route class effect to all other ESAs. That is the assumption used in Table 6.1 to calculate the route-class effect for stand-alone tail end services.

Table 6.1: Route-related effects, preferred RE model

	<i>Inter-capital</i>	<i>Metro.</i>	<i>Regional</i>	<i>Tail end</i>	
Route-class coefficient	0.0000	0.1739	0.3150	n.a.	
<i>Route throughput effect:</i>					
Mean route t'put*	3.8184	4.1685	2.9554	4.5823	
Route t'put coefficient	-0.0197	-0.0197	-0.0197	-0.0197	
Route t'put effect	-0.0751	-0.0819	-0.0581	-0.0901	
<i>ESA throughput effect:</i>					
Mean ESA t'put*	11.7163	9.8426	9.1240	9.0087	
ESA t'put coefficient	0.0303	0.0303	0.0303	0.0303	
ESA t'put effect	0.3546	0.2979	0.2762	0.2727	
<i>Combined effect:</i>					
- Route-class method	0.2796	0.3899	0.5330	<i>Metro.</i> [#]	<i>Regional</i> [#]

Notes: * regulated routes; # Metro or Regional route coefficients assumed to apply to metro ESAs and regional ESAs.

Source: Economic Insights estimation results.

Estimated variance

The estimated variance used for the adjustment shown in equation (6.2) is calculated in (6.3) and the adjustment term is then derived in (6.4).

$$(6.3) \quad \hat{\sigma}^2 = \hat{\sigma}_u^2 + \hat{\sigma}_e^2 = (0.3187)^2 + (0.4291)^2 = 0.2857$$

$$(6.4) \quad \exp\left(\hat{\sigma}^2/2\right) = 1.1536$$

Summary

The combined route-class effect shown in Table 6.1 can be combined with the regression constant to produce the following simplified pricing model for regulated routes:

$$(6.5) \quad M = 1.1536 \cdot \exp\left\{a + 0.4923 \ln C + 0.0950 \ln D - 0.0176 (\ln C)^2 + 0.0068 (\ln D)^2 - 0.0037 (\ln C \ln D) + 0.2434 I\right\}$$

where: a is a combined constant that varies by route-type and using the 'route-class method' shown in Table 6.1:

- Inter-capital routes: $a = 5.0635$
- Metro routes: $a = 5.1738$
- Regional routes: $a = 5.3170$
- Tail ends metro ESAs: $a = 5.1404$
- Tail ends regional ESAs: $a = 5.2815$

6.2.2 Application of Model 3

Model 3 is simpler than Model 2 because ESA throughput and route throughput are omitted. It can initially be expressed as follows:

$$(6.6) \quad M = \exp\{4.9534 + 0.4915 \ln C + 0.1170 \ln D - 0.0175 (\ln C)^2 + 0.0065 (\ln D)^2 - 0.0047 (\ln C \ln D) + 0.2401 I + \delta + \theta\} \cdot \exp(\hat{\sigma}^2/2)$$

Where the notation is the same as for equation (6.2). Again, in the following calculations we assume that the median value of the provider effect is chosen, in which case $\theta = 0$. The combined equation constant and route-class effects in this case are:

- Inter-capital routes: $a = 4.9534$
- Metro routes & tail ends: $a = 4.9534 + 0.2202 = 5.1736$
- Regional routes & tail ends: $a = 4.9534 + 0.3262 = 5.2796$

The estimated variance in this case is:

$$(6.7) \quad \hat{\sigma}^2 = \hat{\sigma}_u^2 + \hat{\sigma}_e^2 = (0.3247)^2 + (0.4286)^2 = 0.2891$$

$$(6.8) \quad \exp(\hat{\sigma}^2/2) = 1.1555$$

The interface-type effect in this model is quite important. Therefore, if a constant value of I were substituted into (6.6) it should be either the sample mean on regulated routes, which is 0.8193, or the sample mean on all routes is 0.7089. The mean for regulated routes would be the most relevant for setting prices on regulated routes.

6.2.3 Rate of change in prices

We omit the derivation of a pricing model from Model 1, but the use of the ‘contract start date’ variable needs to be explained. That variable measures the number of days from 1 January 1960. The mean contract start date in the sample is 18,865.8, and when the coefficient of -0.0000495 is applied to this value, the result is -0.9339, which is approximately equal to the difference between the constants of the models with and without the contract start date variable.

If Model 1 is used for prediction, it is necessary to determine a relevant date for the pricing period over which the price will apply. For example, if the representative date for the application of the price were 1 April 2016 (which is the mid-point between 1 October 2015 and 1 October 2016, and therefore a representative date if the pricing period extended over that term) then the date for a new contract entered into at that time would be 20,545. This means that 1.0170 should be subtracted from the constant (5.6826) to obtain the adjusted constant 4.6656. And at the beginning of the next annual pricing period (1 October 2016 in our example), prices would decrease by 1.8 per cent.

6.3 Tail-end pricing

A ‘tail-end service’ refers to a DTCS service within a single ESA, whether in metropolitan or regional areas (ACCC 2012, p.16). The ACCC (2012b, p. 30) has noted that the vast majority of tail-end services are provided by Telstra in a bundle, with an inter-capital, metropolitan or regional service, and are less than 2 km in length. The ACCC also reported that analysis of tail-end services indicates that tail-end services share some of the same price drivers as other

DTCS services. For tail-end services that are bundled with other services, the charges cover the end-to-end service, and a separate tail-end price is not needed. Our attention in this section is on *stand-alone* tail-end services, which are not bundled with other services. These represent a distinct route classification in the 2014 dataset, and are to be distinguished from the other three route classifications (Inter-capital, Metro and Regional), which all refer to inter-exchange DTCS services (i.e., between different ESAs), perhaps with a bundled tail-end component. Tail-end services are currently regulated.

In the 2012 DTCS FAD, the ACCC used its benchmark pricing model to set stand-alone tail-end prices by assuming that tail-end services have a 2 km standard length, are without protection and have unrestricted speed. The ACCC's assumption that tail-end services have a 2 km standard length for pricing purposes needs to be re-examined, as one stakeholder has strongly contested this approach.

We have calculated indicative tail-end distances using the data for the size of the ESA (in km²) associated with each tail end, and using the following simplifying assumptions to derive the estimates:

- each ESA is circular and the exchanges are located at the centre of the circle
- the average length of a tail end is equal to the average distance of any point within a circle from the centre of the circle (i.e., radius / 2).

Given the area of the ESA (A), the average length of the tail ends associated with that ESA (l)

is estimated using the formula: $l = \frac{1}{2} \sqrt{\frac{A}{\pi}}$, since $l = \frac{r}{2}$ and $r = \sqrt{\frac{A}{\pi}}$.

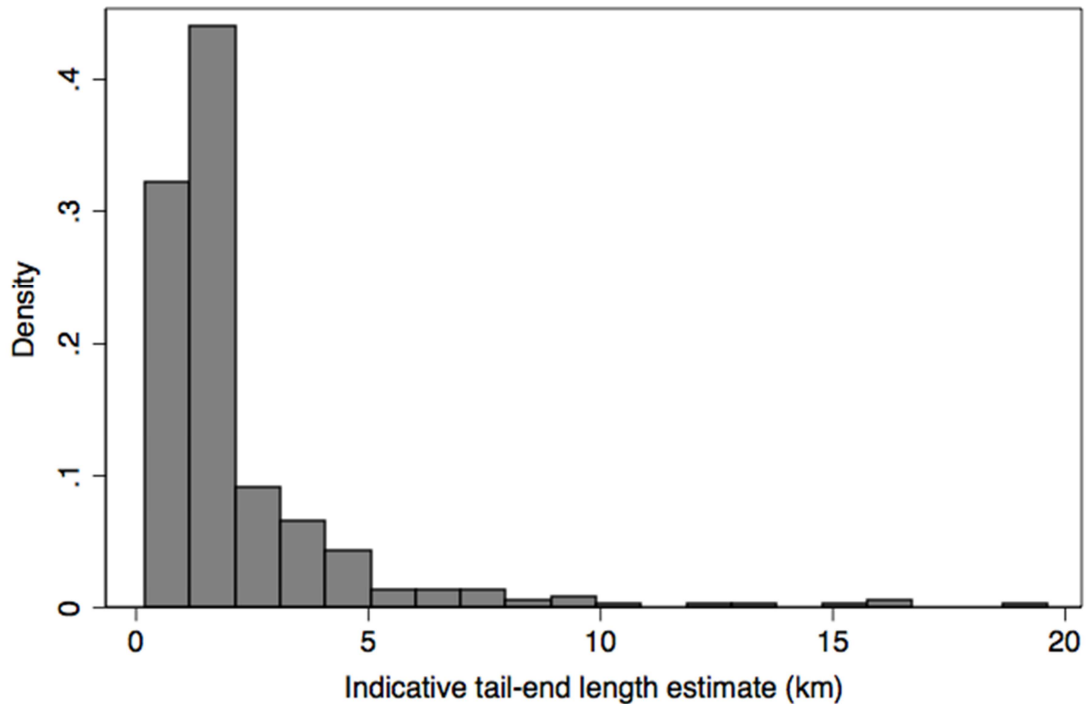
With these estimates we test whether the average tail end length is likely to be in the vicinity of 2 km, and variation of the average tail end length between ESAs. The results are shown in Figure 6.1. The results show that 2 km is a reasonable assumption because it is the modal distance of the estimated average tail-end length per ESA. The assumed typical length of 2 km is also close to the mean value of the indicative estimates.²¹

The indicative tail-end lengths are shorter on average for ESAs located in metropolitan areas than for ESAs located in regional areas. However, quantifying this difference precisely is not feasible because the dataset does not flag whether each A-end and B-end ESA is in a metropolitan or regional area. Only routes are classified as Inter-capital, Metro or Regional, and a route is defined as a pair of ESAs. By definition, Inter-capital and Metro routes have metropolitan ESAs at both ends, so it is possible to identify a set of metropolitan ESAs, but this may not include all of the metropolitan ESAs in the dataset. If we assume that the set of metropolitan ESAs identified in this way is complete, then out of the 409 ESAs with stand-alone tail-end services, 262 are metropolitan and 147 are regional. The average estimated tail-end length on the metropolitan ESAs is 1.30 km, and the average estimated tail-end length on the remaining regional ESAs is 3.74 km.

²¹ The indicative tail-end length, when averaged across all 4,127 stand-alone tail-end contracts, is 1.89 km. When averaged across the 409 ESAs with stand-alone tail-end services, it is 2.18 km.

These observations suggest that the previously assumed average tail-end length of 2 km is reasonable if the price of stand-alone tail-end services is set on a ‘postage stamp’ basis across all metropolitan and regional areas. However, consideration could be given to setting a different tail-end price for metropolitan and regional ESAs, since metro tail-end services are likely to have typical lengths significantly shorter than 2 km, and regional ESAs are likely to have typical lengths significantly longer than 2 km.

Figure 6.1: **Histogram: Average stand-alone tail-end length per ESA (km)**



Source: Economic Insights analysis.

One stakeholder raised the concern that the pricing model cannot be expected to provide an accurate prediction of the cost of stand-alone tail-end prices because there are no stand-alone tail-end services in the data for unregulated routes used to estimate the model. This is a valid issue as a matter of general principle. However, the data sample does include a significant number of services over relatively short distances. As shown in Table 2.1, the data for unregulated routes includes 387 services on Metro routes for distances of less than 1 km, which represents 5.7 per cent of the sample. Whether these services are analogous to stand-alone tail-end services is a question we cannot answer *a priori*, but the distances associated with most tail-end services are not outside the range of values represented in the estimation sample. Absent further information, we would conclude that the econometric model could be used for pricing on stand-alone tail-end routes.

That said, it is recognised that although route-class effects are estimated for the Metro, Regional and (by default) Inter-capital route classes, a corresponding effect is not estimated for the stand-alone tail-end route class. We would suggest that for the ESAs in metropolitan

areas, it may be reasonable to apply the Metro route-class effect, and for ESAs in regional areas the Regional route-class effect could be used.

6.4 Comparison of pricing benchmarks

Table 6.2 presents comparative examples of the predictions from the preferred models, namely Models 2 and 3 shown in Table 5.1. The models used to make these predictions are simplified by assuming:

- the provider effect = 0
- the constants (which include the route-throughput and ESA-throughput effects as shown in Table 6.1) for metro and regional tail-end routes are as shown in section 6.2.1
- tail-end route distance for metro and regional tail-ends is 2 km.

These are compared against those using the DAA model estimated with the 2011 data and with the 2014 data. The DAA model is simplified as indicated above by assuming:

- QOS2, QOS3 and QOS4 = 0
- there is no route-specific effect for tail-end routes, so the same constant applies as for inter-capital routes.
- tail-end route distance = 2 km.

The comparisons are made on the basis of 42 randomly selected observations from regulated routes, as well showing the mean of the predictions for all observations on regulated routes. Summary averages across all regulated routes by route type are also shown.

The following general observations can be made:

- The predictions of price using the DAA model, based on 2011 data, are on average (across all regulated routes) materially higher than those using either of the preferred models presented here, or using the DAA model re-estimated with 2014 data. The DAA 2012 model tends to yield much higher prices for high capacity services than the other models.
- The actual monthly charges on regulated routes are considerably lower than the estimates obtained using the DAA 2012 model.
- The predictions of price using the preferred models are, on average, considerably lower than the prevailing prices on regulated inter-capital and regional routes, but they are similar on average to the prevailing prices for Metro and tail-end routes.

It is also noted that the 2012 DAA model was only used to set prices for DTCS services with capacity of less than 1 Gbps, due to the limitations of 2011 dataset. In the 2014 dataset, on exempt routes, 236 contracts are of 1 Gbps or greater (or 3.5 per cent of the sample), with 174 of these (or 2.6 per cent of the sample) being exactly 1 Gbps and 62 (or 0.9 per cent of the sample) with greater capacities. The average capacity of the 62 contracts with capacity greater than 1 Gbps was 6.07 Gbps. The prediction confidence intervals will get wider the further the capacity value differs from the mean.

Table 6.2: Comparison of predicted costs on regulated routes, selected contracts

Unique ref.	Route type	Actual monthly charge (\$)	Capacity (Mbps)	Distance (km)	Interface type	Protection	Predicted monthly charge (\$)*			
							EI model 2	EI model 3	DAA (2011 data)	DAA (2014 data)
	Metro tail-end		2	2.0	SDH	Yes	375	393	380	480
	Regional		2	6.8	SDH	Yes	512	512	510	522
	Metro		2	12.9	SDH	Yes	480	505	508	494
	Metro		2	9.5	SDH	Yes	462	483	478	478
	Metro tail-end		10	2.0	Other	No	594	622	958	880
	Metro		30	17.5	Other	No	1187	1248	2696	1347
	Regional tail-end		2	2.0	SDH	Yes	432	436	380	480
	Regional		2	96.3	SDH	Yes	736	774	865	699
	Metro tail-end		2	2.0	SDH	Yes	375	393	380	480
	Regional		8	105.5	SDH	Yes	1348	1409	2088	1118
	Metro		2	14.8	SDH	Yes	488	515	522	502
	Metro tail-end		2	2.0	SDH	Yes	375	393	380	480
	Intercapital		40	1374.5	Other	No	2273	2277	8318	2854
	Regional		56	871.5	Other	No	3020	3227	9864	2887
	Regional		2	94.7	SDH	Yes	734	771	862	698
	Metro tail-end		2	2.0	SDH	No	375	393	352	516
	Regional		2	170.3	SDH	Yes	806	856	968	744
	Metro		2	4.6	SDH	Yes	424	436	413	441
	Metro		155	23.6	SDH	Yes	2711	2839	8601	2231
	Metro tail-end		2	2.0	SDH	Yes	375	393	380	480
	Metro		2	53.4	SDH	Yes	583	630	674	578
	Regional		2	101.0	SDH	Yes	741	780	873	703
	Regional tail-end		2	2.0	SDH	Yes	432	436	380	480
	Regional		2	158.6	SDH	Yes	796	845	955	739
	Metro		10	29.6	Other	No	836	890	1510	992
	Regional		2	247.0	SDH	Yes	857	916	1043	775

DTCS Benchmarking Model

Unique ref.	Route type	Actual monthly charge (\$)	Capacity (Mbps)	Distance (km)	Interface type	Protection	Predicted monthly charge (\$)*			
							EI model 2	EI model 3	DAA (2011 data)	DAA (2014 data)
16425	Metro	1215	10	25.6	Other	No	820	871	1467	976
4443	Metro tail-end	243	2	2.0	SDH	Yes	375	393	380	480
12193	Regional	731	2	180.6	SDH	Yes	813	865	980	749
7591	Metro	265	2	4.1	SDH	Yes	418	430	404	436
8144	Metro	1027	4	15.1	SDH	No	667	702	747	680
12072	Regional	1608	2	377.3	SDH	Yes	921	992	1134	812
7128	Metro tail-end	468	2	2.0	Other	Yes	294	309	380	480
7172	Metro tail-end	251	2	2.0	SDH	Yes	375	393	380	480
11115	Regional	2252	2	332.0	SDH	Yes	901	968	1106	801
17851	Regional tail-end	265	2	2.0	SDH	Yes	432	436	380	480
2718	Regional	265	2	2.1	SDH	Yes	450	439	404	459
3652	Metro tail-end	251	2	2.0	SDH	Yes	375	393	380	480
9261	Metro tail-end	243	2	2.0	SDH	Yes	375	393	380	480
13475	Metro tail-end	251	2	2.0	SDH	Yes	375	393	380	480
516	Metro	1420	50	4.6	Other	No	1227	1260	2845	1379
3639	Regional	517	2	176.8	SDH	No	811	861	902	804
Avg regulated routes										
- Intercapital		5077	208	1314	50%	38%	2141	2122	12774	2659
- Metro		836	60	11	75%	75%	820	854	2321	873
- Regional		2819	121	224	85%	81%	1387	1444	5544	1337
- Tail end		442	41	2	88%	88%	522	541	1064	662
- Overall		1399	76	108	82%	80%	917	951	3115	984

Notes: * Metro and regional tail-end lengths assumed to be 2 km. In the EI models, the route-class effects differ between metro and regional tail-ends as discussed in section 6.2.

Source: Economic Insights analysis.

6.5 Allowances for productivity

As explained in this paper and supported by technical experts at the workshop it was considered preferable to use only the 2014 data to estimate a preferred model that could be used for establishing benchmark prices. That approach does not provide any indication of how productivity improvements or more intense competition or other factors could affect prices for the 2015 FAD regulatory period.

It is also apparent that given the large movements in prices for the same service parameters, as seen on competitive routes between 2011 and 2014, any attempt to predict price movements for any period of more than one year would carry a substantial risk of forecasting error.

Given the lack of information to provide a robust empirical basis for specifying a productivity adjustment factor, the ACCC may wish to consider setting prices on an annual basis by updating the statistical model every year. Clearly, however, there is a trade-off between, on the one hand, the cost of regulatory administration and compliance, and on the other hand, the risk of prices becoming inefficient over time due to rapid technological change.

6.6 Setting prices based on the mean or some other percentile

One issue to be considered when using the model to determine benchmark prices on regulated routes is whether to use the mean prediction of the model or make some adjustment to the mean. We do not think there is a persuasive statistical argument to use a prediction other than a mean prediction. Essentially an adjustment to the mean prediction needs to be based on a rationale based on information that is not captured by the model. This could include information about the extent to which predicted prices need to be adjusted down to reflect more scope to reduce costs or up to give greater assurance that revenue was sufficient to cover efficient costs. In this respect, we note that mean values reflect the price that would be charged, on average, if the prices were determined by a market with effective competition. We also note that a mean prediction would be likely to be conservative on the upside if recent price trends continue, particularly if the preferred model excludes the start date variable.

REFERENCES

- AEMC 2013, 'Advice on best practice retail price methodology, Final Report'.
- Arellano, M. & Bonhome, S. 2013, 'Random Effects Quantile Regression'.
- Australian Competition and Consumer Commission (ACCC) 2012, 'Final Access Determination for the Domestic Transmission Capacity Service: Explanatory Statement'.
- Australian Competition and Consumer Commission (ACCC) 2014a, 'Domestic Transmission Capacity Service Final Access Determination Discussion Paper - Primary Prices'.
- Australian Competition and Consumer Commission (ACCC) 2014b, 'Domestic Transmission Capacity Service, Public inquiry into making a final access determination: Position statement on pricing methodology'.
- Australian Competition and Consumer Commission (ACCC) 2014c, 'Domestic Transmission Capacity Service: An ACCC Final Report on the review of the declaration for the Domestic Transmission Capacity Service'.
- Clarke, D. 2014, 'General-to-specific modeling in Stata', *Stata Journal*, vol. 14, no. 4, pp. 895-908.
- Cornwell, C., Schmidt, P. & Sickles, R. 1990, 'Production Frontiers with Cross-Sectional and Time Series Variation in Efficiency Levels', *Journal of econometrics*, vol. 46, pp. 185-200.
- Data Analysis Australia 2012, 'Updated Pricing Model For The Domestic Transmission Capacity Service', Prepared for the ACCC.
- Data Analysis Australia (DAA) 2012, 'Domestic Transmission Capacity Service Price Benchmarking - Pricing Model Development: Consolidated Report', prepared for the Australian Competition and Consumer Commission.
- Economic Insights 2015a, 'Domestic Transmission Capacity Services Benchmarking Model: Draft Report prepared for Australian Competition and Consumer Commission'.
- Economic Insights 2015b, 'Domestic Transmission Capacity Services Benchmarking Model: Workshop Paper prepared for Australian Competition and Consumer Commission', in *Industry Version*.
- Greene, W. H. 2012, *Econometric Analysis*, Seventh (International) edn, Pearson.
- Hausman, J. & Sidak, G. 2013, 'Telecommunications Regulation: Current Approaches with the End in Sight', in *Economic Regulation and Its Reform: What Have We Learned?*, University of Chicago Press, pp. 345-406.
- Said, A. 1992, Modeling Producer Behaviour by Using the Third-Order Translog Cost Function, Concordia University.
- Zellner, A., Keuzenkamp, H. & McAleer, M. (eds) 2004, *Simplicity, Inference and Modeling: Keeping it Sophisticatedly Simple*, Cambridge University Press.
-

ANNEX A: ADDITIONAL DEMAND AND SUPPLY VARIABLES

The ACCC provided additional demand and supply variables that have been considered in constructing a price model.

Demand-related Variables

- NBN POIs – If one of the exchange service areas (ESAs) on the route has a national broadband network (NBN) point of interconnection (POI) then the value for this variable is 1. If both the A-end and the B-end ESAs on the route have NBN POIs then the value is 2. If none of the ESAs have POIs then the value is 0.
- Average number of access seekers – The number of access seekers to the Unconditioned Local Loop Service (ULLS) and Line Sharing Service (LSS) at the A-end and B-end ESAs summed and divided by 2. These two services represent the most basic functions of Telstra's copper network. An access seeker in this case refers to a firm seeking access to the ULLS/LSS in order to provider end user customers with ADSL or voice services. Access seekers lease the copper line from Telstra and provide their own DSLAM in order to provide their own products to end users. This variable is designed to capture derived demand for transmission services.
- Average number of SIOs – The total number of fixed line services in operation (SIOs) as collected via the Telstra Customer Access Network Record Keeping Rule (CAN RKR) at the A-end and B-end ESAs summed and divided by 2.
- SIO density – The average number of SIOs on the route divided by the average size of the ESAs (km²).
- Route throughput (Mbps) – The total contracted capacity for each route in the data set. The sum of the reported capacity for every contract in the data set on the particular route.
- ESA throughput (Mbps) – Unlike the previous metric, this demand metric is non-route specific. The total known contracted DTCS capacity for each A-end and B-end ESA in the data set is calculated as the sum of the reported capacity for every contract in the data set on that particular A-end or B-end ESA.
- Provider-Route throughput (Mbps) – The total known contracted DTCS capacity by service provider for each unique route in the data set. The sum of the reported capacity for every contract in the dataset for the particular provider on the particular route.
- Adjusted number of SIOs (Root Sum of Squares method)²² – The total number of SIOs at each ESA is squared and summed together and then the square root is taken

²² Telstra's public response to the Commission's price terms in the draft final access determination for the Domestic Transmission Capacity Service, 9 March 2012, p.18,

<http://www.accc.gov.au/system/files/Telstra%20Submission%20-%20Draft%20DTCS%20FAD%20-%20Price%20Terms%20-%20March%202012.pdf>

(Root Sum of Squares method). The number of SIOs is an indicator of the demand for retail services, and the demand for transmission services is a derived demand.²³

- Adjusted number of SIOs (Root Sum of Squares elements method) – There are various types of SIOs collected via Telstra CAN RKR e.g. voice only services, ADSL services bundled with voice services, ULLS services, etc. Under this method, each of the different types of SIOs at the A-end is squared separately and summed and added to the sum of the squared elements of the B-end ESA and then the square root is found. The ACCC considers that this method better captures the differences in the units of measurement for the different types of SIOs at the A-end and B-end.²⁴
- Adjusted SIOs (weighted by bandwidth) – The number of voice only SIOs, as recorded in the Telstra CAN RKR is relatively high compared to the other types of SIOs. However, the data rate for voice is low (an average of 0.64 kbps per SIO) compared to the data rate for DSL Broadband (an average of 1088 kbps per SIO) based on a 2008 model by Gibson Quai with variables updated to reflect 2014 usage.²⁵ This method builds on the previous two methods for calculating SIOs as it directly considers the differences in data rates for the different types of SIOs. This method adjusts for the impact of the high number of low data rate POTS SIOs, by weighting each of the four elements that make up total SIOs accounting for the data rate (SIO POTS ONLY, SIO POTS + ADSL, Telstra xDSL no POTS, TOTAL ULLS). A further uplift of 15 per cent is added to account for business services.

Supply-related Variables

- Average number of ‘ESA providers’ - The number of firms with their own transmission infrastructure within 150 meters of a Telstra exchange at the A-end and B-end ESAs summed and divided by 2.
- Number of DTCS transmission providers at A-end or B-end – The number of DTCS transmission service providers identified from the data request information providing services at the A-end ESA or B-end ESA. [REDACTED]
- Number of DTCS transmission providers at A-end or B-end (not top 4) – The number of DTCS transmission service providers identified from the data request information providing services at the A-end ESA or B-end ESA that are not [REDACTED]
- Number of DTCS transmission providers on route (this is referred to as ‘DTCS providers’ in the body of the report – A route is a pair of A-end and B-ends ESAs that

²³ Telstra in their public submission to the ACCC recommend the use of Root Sum of Squares as each type of SIO uses a different level of bandwidth. For instance Telstra xDSL no POTS SIOs use a higher bandwidth than SIO POTS ONLY. Telstra considers that by using the Root Sum of Squares method it has accounted for the different units of measurement for the different types of services at the A-end and B-end to normalise for the differences between each of the different units.

²⁴ $RSSE = \sqrt{(\sum_{i=1}^4 a_i^2 + \sum_{j=1}^4 b_j^2)}$, where a_i is the number of SIOs of type i at the A-end, and analogously for the B-end.

²⁵ ABS 8153.0 - Internet Activity, Australia, June 2014

are identical. This is a count of the number of DTCS transmission providers identified from the data request information providing services on a route.

- Number of DTCS transmission providers on route (not top 4) – The number of DTCS transmission providers identified from the data request information providing services on a route that are not [REDACTED]
- Total unique DTCS transmission services provided from A-end and B-end – The number of DTCS transmission services identified from the data request information being provided from the A-end ESA or B-end ESA on the route.
- Total unique DTCS transmission services provided on route – The number of DTCS transmission services identified from the data request information being provided on the route.

ANNEX B: DATA MANAGEMENT DOCUMENTATION

Item	Initial transformations
All string variables	All string variables were either encoded as numeric variables or dropped from the datasets used in the analysis. Encoding assigns an ordinal integer to each discrete value of the string variable, usually in alphabetical order of those values. String variables that have binary values (eg, “Y” and “N”) were coded as indicator variables (taking values 1 or 0).
Missing & zero values	String values “[blank]” were assigned as missing values. String values “nil” or “Internal Order” were assigned as 0.
Distance	61 missing values for distance (ESA-to-ESA basis) were imputed using estimates derived by estimating the average relationship between ESA-to-ESA distance and other measures of distance in the dataset. Using OLS, (a) ESA-to-ESA distance regressed against Address-to-Address distance, and (b) separately ESA-to-ESA distance was regressed against reported distance. Only data for deregulated routes was used in estimating these equations. The predicted values were then used to impute the missing values for the ESA-to-ESA distance variable, using model (a) whenever the Address-to-Address distance is available, and using model (b) in the remaining cases.
Contract term	186 missing observations on deregulated routes were replaced with the mean contract term on deregulated routes.
Contract start date	Contract start date was converted to Stata date format (which is equivalent to the number of days from 1/1/1960).
Panel variables	The route variable was coded as a numeric variable, and a variable ‘seq’ was created which represents the sequence of observations within each route category, ordered by the ‘unique reference’ variable.
Centering of data	The user-written Stata routine <i>center</i> was used to center data prior to estimation during the specification search.

A validation sample representing 10 per cent of observations on deregulated routes used for testing out-of-sample performance of the models, used during the specification search, was created with the following Stata code:²⁶

```
set seed 19101931
gen randomid = runiform()
sort randomid
gen byte validsample = _n <= 677
```

²⁶ Using Stata version 14, which differs from previous versions (see: <http://www.stata.com/stata14/random-number-generators/>).

ANNEX C: REVIEW OF THE 2012 MODEL

C.1 *Replicating the 2012 model with 2011 data*

The re-estimation of DAA’s 2012 model, using the same 2011 data, is presented in the first model shown in Table C.1. This corresponds to the model reported in DAA (2012), however, we have used the monthly charge as the dependent variable, rather than the annual charge, and the constant is correspondingly smaller. The constant in the original model was 7.6818 and in this model is 5.1969, and these are related as follows: $\exp(7.6818) / \exp(5.1969) = 12$. The other coefficients of the model are unchanged from the original.

A basic goodness-of-fit statistics are shown in the lower part of that table. The R^2 is the squared correlation coefficient between the fitted and actual values of the log monthly charge. The RMSE is the square root of the mean squared residual. MAE is the mean absolute value of the residuals. These statistics can be readily calculated to compare models estimated using different econometric techniques. For some estimation procedures, goodness-of-fit statistics such as adjusted- R^2 , Akaike’s information criterion (AIC) and the Bayesian information criterion (BIC) are not always available. BIC is reported where it is available.

The DAA 2012 model might be interpreted as a cost function (if the prices charged for each service approximate the actual cost of supplying the service) for providing individual transmission services, with capacity and distance being the principal outputs. The elasticities of cost for these two outputs are estimated to be 0.62 and 0.20 respectively, and economies of scale are indicated since the sum of these two elasticities (0.82) is less than one.

Aside from the distance and capacity variables, and whether there is protection, the remaining terms relate to route-class and the ACCC’s quality of service variable, each having some interaction effects. Focussing attention primarily on the main effects, we observe that for the DAA 2012 specification:

- (a) the main effects of the route-class effects are not statistically significant,²⁷ and
- (b) most of the main effects for quality-of-service have an unexpected sign or scale. Recall that *qos* is an ordinal variable, with 1 designed to represent the highest quality and 4 designed to represent the lowest quality. Hence if the variable is an appropriate proxy for quality the coefficients on *qos2* to *qos4* should all be negative, because it should cost less to produce a service of lower quality, and they should be increasing in absolute value. However, these quality measures may be capturing firm specific effects as well. Some firm-specific effects could reflect cost differences associated with exogenous factors but the quality measure is also likely to be a proxy for size and could partially capture residual market power. This issue is considered further in Annex D.

²⁷ In this report we use the term “statistically significant” to refer to a coefficient being insignificantly different from zero at the 5% level of significance, unless we indicate otherwise.

Table C.1: Estimation results, DAA 2012 model & variants (2011 data)*

Predictor	<u>DAA 2012 model</u>		<u>Additional variables</u>		<u>Random effects</u>	
	coefficient	t-stat.	coefficient	t-stat.	coefficient	t-stat.
constant	5.19690	79.01	4.6406	52.96	4.4415	33.03
log capacity	0.62262	58.76	0.7080	32.11	0.7611	36.59
log distance	0.19864	30.87	0.1921	17.18	0.1568	6.91
protection	0.07808	3.27	0.0669	2.77	0.1098	4.69
Route class:						
Metro	-0.08082	-1.47	-0.0788	-1.25	-0.0654	-0.68
Regional	0.05155	0.92	0.1338	2.16	0.2547	2.87
QOS:						
2						
3						
4						
Route class # QOS:						
Metro#2						
Metro#3						
Metro#4						
Regional#2						
Regional#3						
Regional#4						
QOS # log capacity:						
2						
3						
4						
0.5(log capacity) ²			-0.0184	-2.97	-0.0309	-5.32
0.5(log distance) ²			0.0073	1.86	0.0127	1.95
log capacity*log distance			-0.0092	-4.60	-0.0081	-3.87
log route t'put			-0.0089	-2.08	-0.0122	-1.51
log avg. ESA t'put			0.0448	7.30	0.0622	5.73
Interface type			0.1036	3.71	0.0981	3.69
Share of variance due to u _i					0.4545	
Goodness-of-Fit						
R ²	0.8422		0.8464		0.8446 [†]	
BIC	5041.8		4981.8		n.a.	
RMSE	0.4397		0.4339		0.4366 [#]	
MAE	0.3147		0.3087		0.3131 [#]	

Notes: * Dependent variable is log monthly charge; † Square of correlation between log annual charge and fitted log annual charge, where the fitted value does not include the random effect (u_i); # Based on u_i + e_i.

Additional variables

Table C.1 also shows the results of estimating two alternative specifications as variants of the DAA model. The second of the models in Table C.2 ('Additional variables') tests the inclusion of additional predictors, consistent with previous recommendations by Professor Breusch to include second-order terms and demand-related variables. This model includes:

- translog-type higher order and interaction terms for the output variables, capacity and distance
- two additional variables that may be related to demand, supply and/or capacity constraints (route throughput and ESA throughput), and
- an additional quality related variable (interface type).

The purpose of including additional variables is to test whether there are some statistically significant omitted variables, and the implications of including them in the model. We do not suggest that the variables included here are the only candidate additional variables. It is simply one set of possible additional variables. The actual variables included here differ slightly from the variables used in the workshop paper and the draft report (Economic Insights 2015a; b).²⁸ Partly this follows comments received from stakeholders about the meaningfulness of some variables, and partly it is for convenience and consistency.

Joint parameter test statistics for the additional variables are shown in Table C.2, and t-statistics for the individual variables are in Table C.1, and the results can be summarised as follows:

- the second order effects on the two output variables (ie, the squared log of capacity, the squared log of distance, and the interaction between the outputs) are individually significantly different from zero at either a 0.05 or 0.10 level of significance, and they are jointly significant at any level of significance.
- the other additional variables, including total throughput on the route, the average throughput of the ESAs relating to the route and interface type of the service, all in log form, are individually significant at the 0.05 level of significance, and are jointly significant at any level of significance.

Although the additional variables are significant, the improvement in the explanatory power of the model is small, with the R^2 increasing from 0.842 to 0.846. The t-statistics on the main effects of the outputs are reduced, and while the route class variable coefficient for metro remains insignificant, the regional coefficient has increased and is positive and statistically significant.

Table C.2 presents a number of diagnostic tests relating to these models. Misspecification is still present in the ‘additional variables’ model, as indicated by the RESET test and the link test, and while the former tests shows no improvement over the DAA 2012 specification, the link test has improved, although the test statistic of -2.29 still exceeds the critical value of 1.96 in absolute terms. The inclusion of additional variables does not significantly change the other diagnostic test results relating to lack of normality and homoscedasticity of the residuals. These findings suggest that while there appears to be some benefit to including the

²⁸ For example, the Workshop paper tested the log # access seekers, log # SIOs, log ESA throughput and log provider-route throughput as additional variables. Comments at the workshops suggested that log # SIOs and log # access seekers to local loop services were unlikely to have an important influence on DTCS pricing. The route throughput variable used in Table C.1 is an alternative to provider-route throughput. Both variables are strongly correlated with each other.

additional variables in the model, it provides only a small overall improvement.

Random effects

One of the issues of model specification is the likelihood that there are unobserved effects because some of the factors that are relevant to the pricing of different services by different providers on different routes are not available in the dataset. It seems reasonable to expect that route-specific effects are likely to be present, because facilities established at different times in different locations may have different technologies, or their design may be influenced by the topography, urban development or infrastructure in the relevant location or there may be supply/demand imbalances on particular routes. Unobserved effects may be treated as random variables.

The third model shown in Table C.1 tests an alternative stochastic specification, namely the random effects model, to take into account the possible influence of unobserved factors that affect costs differently on different routes. The random effects model allows for a cross-sectional random disturbance across routes, in addition to the usual “white noise” error term, as shown in equation (C.1).

$$(C.1) \quad v_{it} = c_i + u_{it}$$

where v is the combined stochastic term; i indicates the cross-sectional units (here chosen to be routes); t indicates the observations within each cross-sectional unit; c is a random term which varies only between cross-sectional units and represents the unobserved effect; and u is a white noise random term which is independent of c . The random effects model decomposes the disturbance into these two components, thereby estimating a cross-sectional random effect, which is an estimate of the unobserved effect.

The estimation results when the random effects stochastic specification is introduced are shown in the third model in Table C.1, and the related diagnostic tests are in table C.2. The values of the coefficients are altered when random effects are included, but in almost all cases the coefficients take the same sign. The only exception is one of the interaction terms between QOS and route class.

The key findings are:

- The Breusch-Pagan Lagrangian multiplier test supports the relevance of random effects.²⁹
- The RESET test and the link test for misspecification both improved over the other models. The link test statistic of 1.67 is less than the critical value for accepting the null hypothesis that the model is correctly specified. However, the RESET test continues to reject the hypothesis that the model is correctly specified.

²⁹ In Annex D, the Hausman test for random v fixed effects confirms the validity of the random effects specification for the 2014 data.

C.2 Diagnostic Tests

Although the results of some diagnostic tests have been discussed, where they differ between models, many of the diagnostic results are common to all of the models shown in Table C.1, and they are briefly reviewed in this section. Diagnostic tests of the residuals and model specifications are presented in Table C.2.

Table C.2: **Statistical tests, DAA 2012 model & variants (2011 data)**

	<u>DAA 2012 model</u>		<u>Additional variables</u>		<u>Random effects</u>	
	Stat.	P-value*	Stat.	P-value*	Stat.	P-value*
Normality of residuals						
Doornik-Hansen ⁽¹⁾	3176.9	0.0000	3383.0	0.0000	4173.4	0.0000
IQR (% severe outliers) ^{(2)†}	0.58%		0.63%		1.05%	
Influential observations						
Outliers ^{(3)†}	1.54%		1.56%		1.51% ^a	
High leverage ^{(4)†}	5.25%		5.79%		5.79% ^a	
Influential observations ^{(5)†}	3.39%		3.32%		2.83% ^a	
Homoscedasticity						
Breusch-Pagan/Cook-Weisberg ⁽⁶⁾	616.7	0.0000	620.8	0.0000	634.4 ^a	0.0000
Multicollinearity						
# VIF scores > 10	5/17		9/23		9/23	
Misspecification						
RESET ⁽⁷⁾	27.52	0.0000	30.79	0.0000	15.53 ^a	0.0000
Link test ⁽⁸⁾	5.49	0.000	2.28	0.023	1.69 ^a	0.090
Joint parameter tests						
Higher-order output terms ⁽⁹⁾			13.02	0.0000	56.42	0.0000
Additional variables ⁽⁹⁾			23.47	0.0000	46.87	0.0000
Random Effects						
Breusch-Pagan LM test ⁽¹⁰⁾					357.4	0.0000

Note: * Null hypothesis is rejected, as a standard procedure, in these tests, if P-value is less than 0.05. Equivalently, the reported statistic exceeds the critical value for that statistic; † Percentage of $n = 4095$ observations; (1) $\chi^2(2k)$ where $k = 18$ for 1st model, and $k = 24$ for 2nd and 3rd models. (2) Severe outliers represent about 0.0002% of a normal distribution; (3) Studentized residual > 3; (4) Hat value > $3k/n$; (5) Cook's D > $5 \times$ average Cook's D; (6) $\chi^2(1)$; (7) Via powers of the dependent variable, $F(3, n-k-3)$; (8) Absolute value of t -statistic on \hat{h}^2 ; (9) $F(r, n-k-r)$, where $r =$ number of parameters tested, here $r = 3$. (10) $\chi^2(1)$; a Approximate, based on OLS regression of $(y - \hat{u}_i)$ on the predictors.

Tests that primarily relate to the residuals include:

- whether the residuals are normally distributed (primarily needed only for hypothesis tests to be valid and also less relevant where asymptotic tests can be used for large samples);
- the 'influence' of individual observations, including outliers and observations that exert undue influence on the coefficients; and
- homoscedasticity (or constancy of variance) of the residuals.

Tests that relate to the specification of the regression model include:

- identification of high multicollinearity between predictors (which may inflate the estimated variances, affecting the sign and magnitude of the coefficients);
- tests of misspecification in terms of linearity of the functional relationship between the predictors and the dependent variable, the likelihood of omitted variables and the appropriateness of the dependent variable specification.

The tests shown in Table C.2, which relate to the models shown in Table C.1, indicate:

- *Normality of Residuals*: The formal statistical tests clearly reject the null hypothesis of normality of the residuals.³⁰ The distributions of the residuals have fatter tails than the normal distribution. Normality of the distribution of residuals is not essential for unbiased estimates, but is necessary for valid hypothesis testing in small samples. However the sample size is considered to be sufficiently large that non-normality of the residuals is not likely to be an issue of concern.
- *Homoscedasticity*: An important assumption of ordinary least squares (OLS) regression is the homogeneity of variance of the residuals. If the variance of the residuals is non-constant then conventionally measured standard errors of the coefficients may be biased, although White's robust standard errors can be used. The statistical tests suggest that heteroscedasticity is likely to be present in the model.³¹
- *Observations with Undue Influence*: A model may lack robustness if there are a small number of overly influential data points, which may cast doubt on inferences based on the model, or out-of-sample predictions. The *influence* of an observation is the combined effect of being an outlier (where the residual term from the regression is large in absolute value) and having high leverage (where a predictor takes an extreme value relative to its mean). Using standard tests, approximately 60 observations (1.5 per cent of the sample) were identified as outliers and about 20 of these were severe outliers.³² One measure of a data point's overall influence is Cook's D (which measures the effect of deleting that observation from the model). With the first two models shown in Table C.1, 3.3 per cent of the observations are found to be highly influential, and with the third model (random effects) 2.8 per cent were highly influential.³³ These tests suggest that many observations

³⁰ Doornik-Hansen omnibus test and IQR test.

³¹ Breusch-Pagan/Cook-Weisberg test.

³² A method for identifying outliers is to examine the studentized residuals. We describe as an 'outlier' a data point with a studentized residual greater than 3 in absolute value since only about 0.26 per cent of observations drawn from a normal distribution would exceed that value. The IQR test identifies severe outliers based on their distance from the mean measured in multiples of the inter-quartile range. Only 0.0002 per cent of the normal distribution would be severe outliers.

³³ A commonly recommended threshold for an influential data point using the Cook's D statistic is $4/(n - k)$, which is about 0.001 in this sample, although this tends to generate too many points of influence (about 260 in this case, or 6.3 per cent of the sample). Others suggest that a threshold of 1 be used, and the chart of residuals presented by DAA (2012, p.11) suggests they used a threshold of 0.5. These thresholds may be more suitable to small samples. We have used a threshold value of Cook's D equal to: $5 \times \text{mean}(\text{Cook's D})$.

have a high degree of influence. Plots of leverage v squared (normalised) residuals confirmed this. Three of the smaller providers dominated the observations with a high degree of influence.

- *Multicollinearity*: Multicollinearity can become a problem if there is close correlation between predictors, such that a substantial part of the variation of one of the predictors could be explained by a linear function of the other predictors. When there is a high degree of multicollinearity, the coefficient estimates may be poorly identified with large variance and some coefficient estimates may be sensitive to small changes in the data. Multicollinearity is a sampling problem, in the sense that a larger and richer dataset may enable the poorly identified effects to be better identified. One method of detecting high multicollinearity is to calculate variance inflation factors (VIFs) for each explanatory variable. This measures the degree to which the variance of a variable has been inflated because that variable is not orthogonal to other variables. If the VIF value for a variable is greater than 10, this suggests that the variable is close to being a linear combination of other explanatory variables. In the DAA 2012 model, 5 out of 17 variables had VIFs > 10 , whereas in the other two models shown, 9 out of 23 variables had VIFs > 10 . The latter models introduced higher-order terms relating to the outputs, which may explain this difference in the degree of multicollinearity. Although strong multicollinearity can affect the interpretation of the coefficients and even entail reversal of expected sign, multicollinearity is not likely to be a major problem for forecasting if the pattern of multicollinearity in the explanatory variables does not change materially between the data used for estimation and forecast purposes.
- *Misspecification*: Misspecification of a model might be due to adopting an inappropriate functional form (such as assuming a linear relationship between variables that are actually related nonlinearly) or due to omitted variables, for example. Misspecification can result in biased and inefficient estimates. Formal specification tests shown in table C.2 reject the null hypothesis of no misspecification in the DAA 2012 model and in the model with additional variables.³⁴ With the random effects model, the two tests of misspecification have conflicting results and only one of them suggests there is misspecification. One possible misspecification of a model is in the assumption that the dependent variable is a linear function of the predictor variables included in the model when in fact there is a nonlinear relationship. One approach to detecting non-linearity, in models where the variables enter linearly, is to use augmented partial residual plots to visually identify any nonlinearity in the data. These plots suggested that in the DAA 2012 model, linearity may breakdown at the lower and upper values of monthly charges.

C.3 Estimating the DAA specification with 2014 data

Table C.3 presents the results for the same three econometric models as shown in Table C.1, but estimated using the 2014 data. In the 2014 dataset, the interface variable has three types rather than two, giving rise to an additional indicator variable for the third category of interface type. In other respects the model specifications are the same as in Table C.1.

³⁴ Link test for specification of dependent variable and RESET test for omitted variables and functional form.

DAA specification

Some of the key points to note about the re-estimation of the DAA 2012 specification with the 2014 data are:

- The goodness-of-fit is much lower. When applied to the 2014 data, the R^2 is 0.643, compared to 0.842 when the same specification was estimated using the 2011 data. The RMSE is 0.56 and the MAE is 0.40, compared to 0.44 and 0.31 respectively in the corresponding model shown Table C.1.
- There are considerably smaller coefficients on the capacity and distance variables (but they are still highly statistically significant). These two coefficients sum to 0.44, compared to 0.82 when estimated using the 2011 data.
- There is a change in sign for the protection variable, which is inconsistent with the expectation that providing protection involves some additional cost, which may justify a higher price but certainly not a lower price. One interpretation might be that protection tends to be available on routes where it can be more easily provided, but the change from the 2011 data would be difficult to explain, and furthermore, the protection coefficient has a positive sign on both of the other two models shown in Table C.3.
- There is a change in sign for the QOS indicator variables, which is more consistent with expectations, because their coefficients are negative and the absolute values of the coefficients on QOS 3 and 4 are greater than for QOS 2 and QOS 1 (which by implication is zero). However, these main effects cannot be accurately interpreted without taking into account the 9 interaction terms. There are some changes in the signs and magnitudes among the interaction terms between quality and route class and between quality and capacity.

The diagnostic statistics are shown in Table C.4. The same general observations relating to non-normality and heteroscedasticity of the residuals and the significant presence of outliers and observations with a high degree of influence apply to these models when estimated with the 2014 data. Three differences are notable in relation to the DAA 2012 specification:

- Severe outliers are more frequent in the 2014 data. For the DAA 2012 specification, severe outlier residuals represent 0.75 per cent of the sample, compared to 0.58 per cent when estimated with 2011 data.
- There is a higher degree of multicollinearity between the regressors, with 8 out of the 18 coefficients in the DAA 2012 specification having VIF scores are greater than 10 (compared to 5 out of 17 previously).
- The two test statistics for misspecification shown in Table C.4, namely the RESET test and the link test, continue to strongly reject the null hypothesis that the model is correctly specified and have deteriorated for the DAA 2012 specification.

Table C.3: Estimation results, DAA 2012 model & variants (2014 data)*

Predictor	DAA 2012 model		Additional variables		Random effects	
	coefficient	t-stat.	coefficient	t-stat.	coefficient	t-stat.
constant	5.7855	60.88	4.1792	37.63	4.4919	33.05
log capacity	0.3312	39.12	0.5430	39.00	0.5108	37.84
log distance	0.1098	15.88	0.0272	2.37	0.0782	3.21
protection	-0.0725	-4.39	0.0419	2.46	0.0476	2.99
Route class:						
Metro	-0.1762	-2.01	0.1211	1.31	0.0847	0.80
Regional	-0.0499	-0.55	0.2630	2.85	0.1857	1.85
QOS:						
2						
3						
4						
Route class # QOS:						
Metro#2						
Metro#3						
Metro#4						
Regional#2						
Regional#3						
Regional#4						
QOS # log capacity:						
2						
3						
4						
0.5(log capacity) ²			-0.0450	-13.85	-0.0383	-12.50
0.5(log distance) ²			0.0364	9.00	0.0255	3.77
log capacity*log distance			-0.0084	-5.20	-0.0061	-3.78
log route t'put			-0.0326	-8.96	-0.0208	-3.70
log avg. ESA t'put			0.0512	9.18	0.0235	2.83
interface-EoSDH			0.6502	21.00	0.5021	17.07
interface-SDH			0.7268	24.36	0.6486	23.10
Share of variance due to u _i					0.3536	
Goodness-of-Fit						
R-sq	0.6434		0.6886		0.6776 [†]	
BIC	11446.1		10589.4		n.a.	
RMSE	0.5571		0.5206		0.5303 [#]	
MAE	0.3970		0.3719		0.3782 [#]	

Notes: * Dependent variable is log monthly charge; † Square of correlation between log annual charge and fitted log annual charge, where the fitted value does not include the random effect (u_i); # Based on $u_i + e_i$.

Including Additional Explanatory Variables for 2014 Data

The two alternative specifications shown in Table C.3 have considerably better fit than the DAA 2012 specification, whereas previously with the 2011 data, they provided only a marginal improvement in goodness-of-fit. For R^2 the 'Additional variables' model is 0.689, and for the 'Random effects' model is 0.678, which compare favourably to the R^2 of 0.643

for the DAA 2012 specification.

The higher-order output terms and the additional variables are statistically significant, and more strongly so than for the corresponding models in Table C.1. The coefficients on the main effects on log capacity and log distance sum to 0.57 and 0.58 for the ‘Additional variables’ and ‘Random effects’ models respectively, which is greater than the sum of the effects on these variables in the DAA 2012 specification (0.44).

The route-class variables are positive in these alternative models, and the coefficient on the regional route type is greater than for the Metro route type. These implies an ordering of cost between Inter-capital, Metro and Regional route types from lower to higher, which is more meaningful than the coefficients in the DAA 2012 specification. However, this interpretation is again complicated by the interaction terms.

Table C.4: **Statistical tests, DAA 2012 model & variants (2014 data)**

	<u>DAA 2012 model</u>		<u>Additional variables</u>		<u>Random effects</u>	
	Stat.	P-value*	Stat.	P-value*	Stat.	P-value*
<i>Normality of residuals</i>						
Doornik-Hansen ⁽¹⁾	2903.1	0.0000	1798.4	0.0000	2439.1	0.0000
IQR (% severe outliers) ^{(2)†}	0.75%		0.74%		1.11%	
<i>Influential observations</i>						
Outliers ^{(3)†}	1.57%		1.85%		1.63% ^a	
High leverage ^{(4)†}	6.43%		3.74%		4.42% ^a	
Influential observations ^{(5)†}	3.89%		3.52%		3.90% ^a	
<i>Homoscedasticity</i>						
Breusch-Pagan/Cook-Weisberg ⁽⁶⁾	2170.7	0.0000	1666.4	0.0000	1593.2 ^a	0.0000
<i>Multicollinearity</i>						
# VIF scores > 10	8/17		12/24		12/24	
<i>Misspecification</i>						
RESET ⁽⁷⁾	37.56	0.0000	13.38	0.0000	12.80 ^a	0.0000
Link test ⁽⁸⁾	8.60	0.000	1.14	0.256	1.84 ^a	0.066
<i>Joint parameter tests</i>						
Higher-order output terms ⁽⁹⁾			120.39	0.0000	229.09	0.0000
Additional variables ⁽⁹⁾			187.29	0.0000	551.01	0.0000
<i>Random Effects</i>						
Breusch-Pagan LM test ⁽¹⁰⁾					1326.4	0.0000

Note: * Null hypothesis is rejected, as a standard procedure, in these tests, if P-value is less than 0.05. Equivalently, the reported statistic exceeds the critical value for that statistic; † Percentage of $n = 4095$ observations; (1) $\chi^2(2k)$ where $k = 19$ for 1st model, and $k = 25$ for 2nd and 3rd models. (2) Severe outliers represent about 0.0002% of a normal distribution; (3) Studentized residual > 3; (4) Hat value > $3k/n$; (5) Cook’s D > $5 \times$ average Cook’s D; (6) $\chi^2(1)$; (7) Via powers of the dependent variable, $F(3, n-k-3)$; (8) Absolute value of t -statistic on hat^2 ; (9) $F(r, n-k-r)$, where $r =$ number of parameters tested, here $r = 3$. (10) $\chi^2(1)$; ^a Approximate, based on OLS regression of $(y - \hat{u}_i)$ on the predictors.

The diagnostic statistics in Table C.4 show that the observations relating to non-normally distributed and heteroscedastic residuals, and the relatively large number of influential data points, apply equally to the alternative models. The high degree of multicollinearity remains a

particular concern. But these two models do perform better in terms of the misspecification tests. Although the RESET test continues to reject the hypothesis of a correctly specified model, and suggests there are important omitted variables, the link test is satisfied for both of these models, which suggests that the specification is a considerable improvement over the DAA 2012 specification with the 2014 dataset.

ANNEX D: ECONOMETRIC SPECIFICATION SEARCH

The purpose of this appendix is to document the main steps in the process of developing a preferred model. Most of the process is documented in detail in the workshop paper and the draft report (Economic Insights 2015a; b), which report the models estimated at each step of the process. It is not necessary or desirable to reproduce that amount of detail here.

Estimation methods

The general-to-specific methodology and other aspects of the methodology are explained in chapter 5.

First stage results

The first round of the analysis was to estimate the general model described in equation (5.3). In keeping with the general-to-specific modelling approach, the models estimated in the first round include all of the feasible conditioning variables from the dataset provided by the ACCC, and including the higher-order and interactions implied by equation (5.3).

Six models are initially tested to resolve the issue of the appropriate estimation methods. These are:

- (1) Ordinary least squares (OLS)
- (2) Quantile regression
- (3) Robust regression
- (4) Fixed effects model (with route-specific effects)
- (5) Random effects model (with route-specific effects)
- (6) Quantile regression (with data transformed for random effects estimation).

Because extreme values are potentially a problem in the OLS model, two alternative estimation methods have been used, namely quantile regression and robust regression. Quantile regression at the median is one method of obtaining a central representative plane through the data, but is based on least absolute deviation (rather than minimum squared deviation) and is therefore less affected by extreme values. Robust regression is method of regression which assigns different weights to observations, with less weight to outliers.

The results of estimating the general model using these six estimation methods are shown in table D.1. Issues relating to the statistical significance of variables are dealt with in the second stage, which focuses on simplifying the general model. The initial stage is focussed on the overall performance of the different estimation methods.

With regard to comparing the OLS, quantile and robust regression models (ie, the first three shown in table D.1) the main finding is that the fit of the robust regression model is significantly worse than the other two. The RMSE of the OLS model will always be smaller *within sample* than for either quantile or robust regression methods because the OLS technique is to minimise the RMSE. Conversely, since the quantile regression model minimises MAE *within sample*, it will perform better on this measure than OLS. These

conditions need not hold out-of-sample.

Table D.1 shows that:

- The RMSE of the OLS model within sample is 0.4726, which is only slightly lower than 0.4902 for the quantile regression, but considerably lower than the RMSE of 0.5298 for the robust regression model. Similarly, the RMSE of the OLS model out-of-sample (using the validation sample) is 0.4789 compared to 0.4945 for the quantile regression model and 0.5276 for the robust regression model.
- For the OLS model within sample the MAE is 0.3323, which is slightly higher than the quantile regression's MAE of 0.3221. The within sample MAE for the robust regression of 0.3285 is higher than for quantile regression, but less than OLS. Out-of-sample, the MAE of the quantile regression model is 0.3340, compared to 0.3379 for OLS and 0.3419 for robust regression. So the out-of-sample MAE of the robust regression model is higher than for the other two methods.

These comparisons of goodness-of-fit measures indicate that in this context, the quantile regression method is preferred to the robust regression method as a means of limiting the influence of severe outliers. The robust regression method was not used in further analysis. Although the OLS method has slightly better fit in terms of the RMSE and R^2 , the quantile regression outperforms it in terms of the MAE. These two estimation methods were retained in the second stage of the analysis.

Fixed or random effects?

This section discusses tests relevant to the fixed and random variable models in relation to the nature of unobserved route-specific effects which influence the cost of supply. These models allow for such effects in different ways. Table D.1 shows the fixed effects model using centred data. A number of variables are omitted in the fixed effects estimation due to multicollinearity. The distance variables are insignificant in this model, as the fixed effects effectively capture the effects of distance. This reason alone should be sufficient to reject this model. Table D.1 also shows the random effects model using centred data.

Table D.2 shows test statistics relating to the route-specific effects in these models. The hypothesis that all of the fixed effects (u_i) are equal to zero can be tested using an F-test. In this case $F(1508, 4540) = 2.64$. This exceeds the critical F value at a 0.01 level of significance of about 1.1, which indicates an unobserved route-specific effect may be present.

We can test the hypothesis that the route-specific effects are adequately modelled as random effects using the Hausman test (see Table D.2). This is a test of random effects versus fixed effects. This effectively tests whether the unobserved route specific effect is correlated with the other variables in the model, which would mean that all coefficients in the model would be subject to bias. Because in the fixed effects model many of the variables are dropped, the number of coefficients being tested differs between the two models, raising a potential problem in applying the test. Nevertheless, Stata provides the following test statistic: $\chi^2(2) = 0.10$, and $\text{Prob} > \chi^2 = 0.9519$. This indicates that the hypothesis of random effects is not rejected, and with a high level of confidence.

The hypothesis that all of the random effects (u_i) are equal to zero is tested using the Breusch-

Pagan LM test for random effects. In this case the test statistic is $\text{chibar}^2(1) = 818.26$, and $\text{Prob} > \text{chibar}^2 = 0.0000$. This indicates that the hypothesis of no random effects is strongly rejected. These tests support the use of the random effects model in preference to either the fixed effects model or OLS.

Quantile regression with random effects

Table D.1 also shows an estimation of the quantile regression method using data transformed for random effects estimation). This is carried out with the aid of the Stata tool *xtdata*, which produces a transformed dataset of the regression variables suitable for random-effects or fixed-effects estimation. This is designed to aid specification search. Once the data are transformed, a model can be estimated using other Stata commands such as *regress* (for OLS) or *qreg* (for quantile regression). To construct the random effects transformed data, it is necessary to specify the ratio of the standard errors of the two stochastic components, that is σ_u/σ_e , and this was obtained from the fifth model shown in table D.1. The transformation of the data creates a new variable ‘constant’ which is to be included as a regressor, while the intercept must be suppressed. However, the quantile regression procedure in Stata does not allow the intercept to be suppressed. As a result there is a “constant adjustment” shown in Table 5.3, which should represent the difference between the sample mean and the sample median.

Since the random effects quantile regression model cannot be implemented properly at present in Stata, it is not a preferred model in this analysis. (For detail on the random effects quantile regression model see: Arellano & Bonhome 2013).

In its submission to the draft report, one stakeholder indicated that a procedure for estimating a robust regression model with random effects is available in the open source program R. However, the results they reported seemed to show very little difference between the performance of the random effects (RE) and Robust RE models based on RMSE and MAE. The RE model did slightly better on RMSE: 0.504 (RE model) compared to 0.507 (Robust RE). On the other hand, Robust RE did slightly better on MAE: 0.352 (RE model) and 0.348 (Robust RE). These results do not provide strong grounds for using the more complex Robust RE method.

Table D.1: Fixed & random effects models (2014 data)

<i>Predictor</i>	<u>1. OLS</u>		<u>2. Quantile</u> <u>(median)</u>		<u>3. Robust Reg</u>		<u>4. Fixed effects</u>		<u>5. Random effects</u>		<u>6. Quantile with RE</u>	
	<i>Coeff.</i>	<i>t-stat.</i>	<i>Coeff.</i>	<i>t-stat.</i>	<i>Coeff.</i>	<i>t-stat.</i>	<i>Coeff.</i>	<i>t-stat.</i>	<i>Coeff.</i>	<i>t-stat.</i>	<i>Coeff.</i>	<i>t-stat.</i>
Constant	-0.0002	-0.04	-0.0016	-0.28	0.0058	1.32	0.0321	0.04	0.0034	0.24	0.0446	0.53
(constant adjustment) [†]											-0.0270	-0.82
log capacity	2.5184	24.76	2.4971	25.45	2.3353	31.60	2.4302	20.23	2.3438	23.48	2.4949	27.10
log distance	0.2741	2.33	0.5214	4.60	0.8470	9.93	-44.2828	-0.05	0.6172	4.31	0.6409	4.76
0.5(log capacity) ²	-0.0549	-2.86	0.0063	0.34	0.0343	2.46	-0.0355	-1.76	-0.0398	-2.19	-0.0058	-0.34
0.5(log distance) ²	-0.0175	-0.95	0.0514	2.89	0.0378	2.82	16.4144	0.05	-0.0405	-1.08	0.0310	0.86
(log capacity)(log distance)	-0.0263	-3.28	-0.0466	-6.02	-0.0661	-11.34	-0.0171	-2.00	-0.0258	-3.32	-0.0386	-5.39
(1/6)(log capacity) ³	0.0111	2.48	-0.0061	-1.42	-0.0160	-4.93	0.0052	1.07	0.0063	1.48	-0.0023	-0.58
(1/6)(log distance) ³	0.0187	3.02	-0.0066	-1.10	-0.0057	-1.27	omitted	.	0.0176	1.69	-0.0065	-0.65
0.5(log capacity) ² (log distance)	0.0005	0.32	0.0066	4.75	0.0117	11.21	0.0004	0.25	-0.0005	-0.33	0.0029	2.25
0.5(log capacity)(log distance) ²	0.0044	2.35	0.0020	1.11	0.0011	0.85	-0.0001	-0.04	0.0039	2.06	0.0023	1.31
log # DTCS providers	-0.0058	-0.13	0.0320	0.77	0.0496	1.58	omitted	.	-0.0312	-0.45	-0.0181	-0.28
log route t'put	-0.0037	-0.30	0.0005	0.04	0.0009	0.10	omitted	.	0.0093	0.57	0.0224	1.45
log provider route t'put	0.0005	0.04	0.0048	0.48	-0.0090	-1.20	-0.0242	-2.35	-0.0156	-1.56	-0.0114	-1.24
log ESA t'put	0.0870	8.48	0.0563	5.68	0.0701	9.39	omitted	.	0.0879	6.18	0.0658	4.93
log # DTCS services	-0.0050	-0.25	0.0066	0.34	-0.0009	-0.06	omitted	.	0.0219	0.70	0.0248	0.62
contract start date	0.0001	9.62	0.0001	9.12	0.0001	12.02	0.0001	8.41	0.0001	8.72	0.0001	9.10
contract term	0.0075	8.27	0.0116	13.34	0.0115	17.57	0.0041	4.29	0.0060	6.85	0.0087	10.74
Protection	-0.0967	-2.48	-0.0602	-1.60	-0.1487	-5.25	-0.1800	-4.64	-0.1177	-3.23	-0.0234	-0.70
(log DTCS providers)(log capacity)	0.0034	0.31	-0.0115	-1.09	-0.0308	-3.86	0.0503	3.50	0.0195	1.62	-0.0126	-1.12
(log DTCS providers)(log distance)	0.0211	1.74	0.0065	0.55	0.0190	2.16	omitted	.	0.0104	0.59	0.0207	1.24
(log route t'put)(log capacity)	-0.0053	-1.47	-0.0036	-1.03	-0.0052	-1.98	-0.0159	-3.63	-0.0077	-2.10	-0.0020	-0.60
(log route t'put)(log distance)	0.0013	0.47	-0.0034	-1.27	-0.0054	-2.72	omitted	.	-0.0031	-0.89	-0.0120	-3.77
(log provider-route t'put)(log capacity)	-0.0055	-1.49	-0.0020	-0.55	-0.0021	-0.77	0.0027	0.75	-0.0010	-0.30	-0.0011	-0.36
(log provider-route t'put)(log distance)	0.0008	0.29	0.0003	0.12	0.0061	3.07	0.0027	1.03	0.0015	0.59	0.0031	1.32
(log ESA t'put)(log capacity)	-0.0196	-5.2	-0.0196	-5.38	-0.0214	-7.78	-0.0202	-3.34	-0.0188	-4.57	-0.0213	-5.61
(log ESA t'put)(log distance)	-0.0132	-3.33	-0.0067	-1.76	-0.0102	-3.53	omitted	.	-0.0175	-3.68	-0.0089	-1.98

DTCS Benchmarking Model

<i>Predictor</i>	<u>1. OLS</u>		<u>2. Quantile</u> <u>(median)</u>		<u>3. Robust Reg</u>		<u>4. Fixed effects</u>		<u>5. Random effects</u>		<u>6. Quantile with RE</u>	
	<i>Coeff.</i>	<i>t-stat.</i>	<i>Coeff.</i>	<i>t-stat.</i>	<i>Coeff.</i>	<i>t-stat.</i>	<i>Coeff.</i>	<i>t-stat.</i>	<i>Coeff.</i>	<i>t-stat.</i>	<i>Coeff.</i>	<i>t-stat.</i>
(log DTCS services)(log capacity)	0.0174	3.34	0.0122	2.42	0.0166	4.38	0.0059	0.90	0.0061	1.09	0.0110	2.11
(log DTCS services)(log distance)	-0.0355	-5.66	-0.0167	-2.75	-0.0127	-2.78	omitted	.	-0.0225	-2.38	-0.0137	-1.50
(contract start date)(log capacity)	-0.0001	-20.29	-0.0001	-21.04	-0.0001	-26.78	-0.0001	-16.99	-0.0001	-18.74	-0.0001	-21.88
(contract start date)(log distance)	0.0000	-2.7	0.0000	-3.27	0.0000	-5.28	0.0000	-2.60	0.0000	-3.33	0.0000	-2.05
(contract term)(log capacity)	-0.0020	-10.6	-0.0037	-19.91	-0.0041	-29.51	-0.0019	-8.64	-0.0017	-9.19	-0.0031	-17.57
(contract term)(log distance)	0.0004	1.83	0.0007	3.13	-0.0001	-0.72	0.0011	4.39	0.0006	2.59	0.0011	5.32
(protection)(log capacity)	0.0220	2.49	0.0199	2.33	0.0390	6.08	0.0503	5.54	0.0357	4.27	0.0157	2.04
(protection)(log distance)	0.0260	3.63	0.0342	4.94	0.0229	4.41	0.0377	4.42	0.0173	2.45	0.0248	3.83
route class 2 (Metro)	-0.7334	-1.68	0.9130	2.17	1.5993	5.04	omitted	.	0.3765	0.63	1.8377	3.28
route class 3 (Regional)	-0.4422	-1.07	1.1955	3.01	1.7963	6.00	omitted	.	0.5480	1.00	2.0348	3.93
(route class 2)(log capacity)	-0.0679	-2.75	-0.0973	-4.08	-0.0906	-5.04	-0.1315	-3.89	-0.0975	-3.75	-0.1336	-5.55
(route class 2)(log distance)	0.0161	0.80	-0.0224	-1.15	-0.0667	-4.55	-0.0671	-2.19	-0.0333	-1.56	-0.0593	-3.00
(route class 3)(log capacity)	0.2310	2.92	-0.0301	-0.40	-0.1864	-3.25	omitted	.	0.0231	0.21	-0.1784	-1.76
(route class 3)(log distance)	0.0921	1.41	-0.1575	-2.50	-0.2265	-4.77	omitted	.	-0.0212	-0.25	-0.2626	-3.23
QOS 2												
QOS 3												
QOS 4												
(QOS 2)(log capacity)												
(QOS 3)(log capacity)												
(QOS 4)(log capacity)												
(QOS 2)(log distance)												
(QOS 3)(log distance)												
(QOS 4)(log distance)												
interface-type 2 (EoSDH)												
interface-type 3 (SDH)	0.4719	7.37	0.5348	8.66	0.4112	8.84	0.2924	4.71	0.3876	6.48	0.5425	9.84
(interface-type 2)(log capacity)	0.0222	1.71	0.0258	2.06	0.0702	7.44	0.0204	1.46	0.0323	2.55	0.0207	1.78
(interface-type 3)(log capacity)	-0.0210	-1.67	-0.0014	-0.11	0.0981	10.69	0.0146	1.13	0.0011	0.10	-0.0049	-0.44
(interface-type 2)(log distance)	0.0524	3.94	0.0274	2.14	-0.0219	-2.27	0.0567	3.76	0.0513	3.97	0.0239	2.02

DTCS Benchmarking Model

<i>Predictor</i>	<u>1. OLS</u>		<u>2. Quantile (median)</u>		<u>3. Robust Reg</u>		<u>4. Fixed effects</u>		<u>5. Random effects</u>		<u>6. Quantile with RE</u>	
	<i>Coeff.</i>	<i>t-stat.</i>	<i>Coeff.</i>	<i>t-stat.</i>	<i>Coeff.</i>	<i>t-stat.</i>	<i>Coeff.</i>	<i>t-stat.</i>	<i>Coeff.</i>	<i>t-stat.</i>	<i>Coeff.</i>	<i>t-stat.</i>
(interface-type 3)(log distance)	0.0477	3.95	0.0135	1.16	-0.0447	-5.09	0.0425	3.03	0.0380	3.21	0.0020	0.18
$\sigma(u)$									0.2605			
$\sigma(e)$									0.4067			
Goodness-of-fit												
R ² *	0.7427		0.7236		0.6808	0.7427	0.0980		0.7359		0.6945	
BIC	8633.8		.		.	8633.8	4901.7		.		.	
RMSE (in sample) [#]	0.4726		0.4902		0.5298	0.4726	0.3511		0.3885#	(0.478)	0.4204	
RMSE (validation) [#]	0.4789		0.4945		0.5276	0.4789	.		.	(0.491)	0.4234	
MAE (in sample)**	0.3323		0.3221		0.3285	0.3323	0.2228		0.2682#	(0.335)	0.2809	
MAE (validation)**	0.3379		0.3340		0.3419	0.3379	.		.	(0.344)	0.2919	

Source: Economic Insights estimation results.

Notes: * Squared correlation between fitted and actual dependent var. (where fitted value does not include u_i) over all data, both within sample and out-of-sample. # Based on e only (RMSE & MAE based on $e + u_i$ shown in brackets). ** Mean absolute error. † with the random effects transformed data, a variable ‘constant’ is created, and this is used as a regressor with the constant suppressed. However, excluding the constant is not an option with quantile regression in Stata. The ‘constant adjustment’ is the resulting intercept, which should represent the difference between the sample mean and the sample median.

Table D.2: Tests statistics, fixed & random effects

	<u>FE model</u>		<u>RE model</u>	
	Stat.	P-value*	Stat.	P-value*
F-test (significance of fixed effects) [F(1508,4540)]	2.64	0.0000		
Hausman test (random v fixed effects) [chi ² (2)]			0.10	0.9519
Breusch-Pagan (significance of random effects) (chibar ² (1))			818.26	0.0000

Source: Economic Insights analysis.

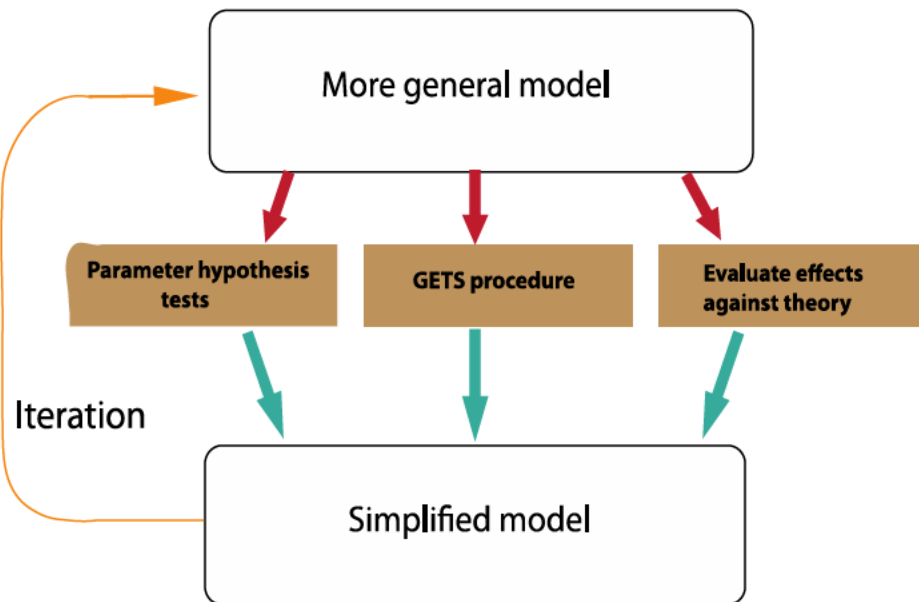
Summary conclusion

The first stage of the analysis narrowed down the preferred estimation methods to quantile (median) regression and the random effects model, while OLS was also retained in case a simpler model was preferred.

Second stage: Simplifying the model

The previous two sections considered the estimation methods in the context of a general model which included all candidate variables and their transformations and interactions within the third-order translog specification. The second stage involved removing unnecessary variables to obtain a more parsimonious model. This section summarises the process of identifying the variables to remove from the model. The methodology is described by the iterative procedure shown in Figure D.1.

Figure D.1: Iterative procedure for model simplification



Source: Economic Insights analysis.

Figure D.1 shows three considerations in the simplification process:

- Firstly, joint hypothesis tests are calculated for each conditioning variable and the interaction terms associated with that variable, and for the third-order terms involving only outputs. These joint parameter tests are Wald tests of the null hypothesis that the ‘true’ value of each parameter is zero. If the p-value exceeds the chosen level of significance, then the null hypothesis cannot be rejected. In relation to the conditioning variables, we adopted a significance level of 0.01 for these tests (and higher significance levels for interaction terms involving only the outputs).
- The second method is to use a general-to-specific (GETS) model selection algorithm. We used the *genspec* user-written routine for Stata (Clarke 2014). This is applied to both the centred data and the data transformed for random effects estimation (using *xtdata*). The GETS procedure drops the variables that contribute least to the model to derive a more parsimonious model. The GETS process was undertaken in two steps. For the first step started the starting point was the general models shown in Table D.1 (for the OLS and RE models only). In the first step a limiting t-statistic of 1.96 was used. The second step commenced from a more parsimonious model and used a higher level of significance as the criterion for excluding variables. The reasons for undertaking the GETS analysis in two steps are as follows:
 - This procedure is indifferent to whether an interaction term or a main effect is excluded, but we give interaction terms associated with a variable higher priority for exclusion than the main effect for that variable. This is the purpose of the joint parameter tests. The main effect is excluded only if, after the exclusion of the interaction terms, it remains insignificant. Similarly, rather than retaining a mix of selected 1st, 2nd and 3rd-order output effects, our preference is to retain the 1st order output effects and give priority to the 2nd order effects over the 3rd order output effects. If the 3rd-order output effects are insignificant, they are jointly excluded.
 - The outcome of the GETS procedure can be sensitive to the initial set of included variables, and since the order in which we exclude variables from the model differs from the order used in the GETS procedure (for the reasons given) it is desirable to undertake this process in two steps, since other criteria are also being considered.
- As indicated in Figure D.1, some regard is also had to the meaningfulness of coefficient signs in terms of their economic interpretation, including in regard to interaction terms. These considerations were used mainly in the final stage of the analysis to achieve greater parsimony than the statistical tests alone would indicate.

Applying this process, the following simplifications were made:

- The main effects and interaction terms for log # DTCS providers, log # DTCS services and log provider-route throughput were found to be insignificant at the chosen level of significance and all of these terms were removed. There was a high degree of correlation between: (a) log DTCS providers; (b) log route throughput; (c) log provider-route throughput; and (d) log DTCS services. Therefore, the removal of some of these variables altered the statistical significance of those retained. Comparative test were carried out to

confirm that the appropriate variables were removed. Log route throughput was the only one of these variables retained.

- The interaction terms for log route throughput were found to be insignificant and were removed.
- The protection variable and its interaction terms were removed in accordance with the results of the GETS procedure.
- The GETS procedure suggested that three of the four interaction terms between route class and the outputs should be dropped, and to aid simplification, all of these interaction terms were removed.
- After simplifying the model, the main effect on interface type 2 (EoSDH) became insignificant and the interaction terms for this variable were jointly insignificant, and all of these terms were excluded from the model. The main effect and interaction terms in interface type 3 (SDH) remained statistically significant, and were retained.
- The four third-order output terms were found to be jointly insignificant and all of these effects were removed, so the model reduced to a conventional translog cost function rather than a third-order translog cost function.
- For the variables contract start date and contract term, the interaction terms with log distance were not significant and were removed.

The following findings relate to the ACCC's quality-of-service categories (QOS), shown in Table D.1. These categories represent an ordering, so that QOS 3 is designed to represent a lower quality than QOS 2, and QOS 4 is designed to represent lower quality than QOS 3. This implies that the coefficients should all be negative and increasing in absolute value. Hence if the variable is an appropriate proxy for quality the coefficients on QOS2 to QOS4 should all be negative, because it should cost less to produce a service of lower quality, and they should be increasing in absolute value. However, the coefficients applying to QOS 2 to QOS 4 were typically positive rather than negative, as expected. QOS 4 was always strongly significant, but the coefficients for QOS 2 and QOS 3 were only significantly different from zero in the quantile (median) model. Also, hypothesis tests showed that the main effects for QOS 2 and QOS 3 were not significantly different from each other.

Furthermore, the coefficients for the interaction of QOS 2 and QOS 3 with log distance were not significantly different from each other (although their interaction terms with log capacity were significantly different). In summary, these tests indicate that there was not much overall difference between the effects applying to QOS 2 and QOS 3 in the model, and nor was there much difference between these effects and those implicitly applying to QOS 1 (subsumed in the intercept or other coefficients in the model). This suggests an approach in which the providers could be alternatively categorised into the larger providers and the smaller providers. To this end we defined a new grouping of providers as Tier 1 and Tier 2: with Tier 1 providers being the four with the largest number of contracts in the dataset as a whole (including regulated and deregulated routes), namely [REDACTED] [REDACTED] [REDACTED] [REDACTED]. The

remaining providers were grouped into Tier 2, and include [REDACTED]. In the preferred model at the draft report stage, the QOS variables were replaced with the Tier 2 variable and its associated interactions with the output variables.

The models that were presented as preferred models at the draft report stage are shown in Table D.3. The OLS model is included only for comparison, and the quantile (median) and random effects models were preferred at that stage of the analysis.

Table D.3: OLS, quantile (median) & random effects models (centred 2015 data)

<i>Predictor</i>	<u>OLS</u>		<u>Quantile (median)</u>		<u>Random Effects</u>	
	<i>Coeff.</i>	<i>t-stat.</i>	<i>Coeff.</i>	<i>t-stat.</i>	<i>Coeff.</i>	<i>t-stat.</i>
constant	-0.0013	-0.20	0.0098	1.46	0.0013	0.08
log capacity	2.1895	27.00	2.3470	27.82	2.0451	25.10
log distance	0.2574	5.97	0.3903	8.70	0.3059	5.34
0.5(log capacity) ²	-0.0209	-5.31	0.0038	0.93	-0.0232	-6.31
0.5(log distance) ²	0.0258	6.58	0.0177	4.32	0.0206	2.99
(log capacity)(log distance)	0.0058	3.13	0.0024	1.25	0.0039	2.16
log route t'put	-0.0314	-8.61	-0.0128	-3.37	-0.0267	-4.54
log ESA t'put	0.1102	13.61	0.1159	13.77	0.1070	8.13
contract start date	0.0001	9.98	0.0001	9.22	0.0001	7.84
contract term	0.0111	13.34	0.0100	11.48	0.0092	11.33
(log ESA t'put)(log capacity)	-0.0170	-5.46	-0.0216	-6.69	-0.0162	-4.64
(log ESA t'put)(log distance)	-0.0217	-6.49	-0.0288	-8.27	-0.0231	-5.07
(contract start date)(log capacity)	-0.0001	-19.66	-0.0001	-20.92	-0.0001	-18.44
(contract term)(log capacity)	-0.0024	-18.62	-0.0027	-19.71	-0.0019	-13.83
route class 2 (Metro)	0.1695	3.14	0.1531	2.73	0.1314	1.69
route class 3 (Regional)	0.2950	6.37	0.2331	4.84	0.3495	5.70
Tier 2	-0.9715	-15.22	-0.6825	-10.27	-0.8687	-14.38
(Tier 2)(log capacity)	0.1061	9.06	0.0318	2.61	0.0814	7.40
(Tier 2)(log distance)	0.0500	5.58	0.0500	5.37	0.0468	5.26
interface-type 3 (SDH)	0.2148	6.40	0.1863	5.33	0.2192	6.87
(interface-type 3)(log capacity)	-0.0337	-3.90	-0.0371	-4.13	-0.0279	-3.43
(interface-type 3)(log distance)	0.0244	3.18	0.0448	5.62	0.0252	3.30
$\sigma(u)$	0.3334	.
$\sigma(e)$	0.4241	.
Goodness-of-fit						
R ² *	0.7106	.	0.7018	.	0.7055	.
BIC	9055.2
RMSE (in sample)	0.5010	.	0.5086	.	0.5051**	.
RMSE (validation)	0.5104	.	0.5186	.	0.5166**	.
MAE (in sample)	0.3497	.	0.3443	.	0.3492**	.
MAE (validation)	0.3519	.	0.3497	.	0.3551**	.

Source: Economic Insights estimation results.

Notes: * Squared correlation between fitted and actual dependent; ** Based on ue.

included in the dataset.

Third stage: Further simplification after the draft report

Comments were made on the foregoing models in submissions to the draft report from the stakeholders and technical experts. The following suggestions were made in relation to the modelling method and specification:

- The quantile regression model be dropped in favour of the RE model. The inclusion of random effects in the model was supported. Although the estimates obtained using the quantile model were similar to the RE model, one expert found the quantile regression model to be unstable (i.e. requiring a large number of iterations to converge and with non-convergence with some sets of predictors) and the rationale of using median is much weaker than that for using the mean. We accept this recommendation, and focus on the random effects model.
- Use the maximum likelihood random effects (ML-RE) option, rather than conventional random effects. The final model(s) be estimated using uncentered data to avoid the need to calculate the intercept, and that the entire data sample of exempt routes contracts (i.e. including the validation sample) should be used in the final analysis. These approaches are all adopted.
- Use a stochastic frontier analysis (SFA) model, rather than a random effects model. However, the cross-sectional SFA model discussed and apparently tested by the expert who made this recommendation did not appear to be consistent with the economic arguments it put forward, which tended to suggest that due to widespread bundling practices, some of the providers may retain some degree of market power. We have addressed this argument in a different way to that proposed by the expert, which permits the well-indicated random route-specific effects to be retained in the model, while introducing provider-specific fixed effects. This can be interpreted as a frontier model of the kind developed by Cornwell et al. (1990), if the differences between u_i (the fixed effect for provider i) and u_{min} (the minimum fixed effect) are treated as distances from a frontier. Alternatively, the provider-specific effects might be considered to reflect a number of factors such as quality differences, inefficiency, differences in market power, or data quality (for very small providers). This approach is an alternative to the ‘Tier 2’ variable used at the draft report stage.
- Consider whether to drop certain interaction terms on conditioning variables in order to simplify the model further, including:
 - the interaction terms relating to interface type
 - the interaction terms between ESA throughput and log capacity and log distance
 - the interaction terms between the outputs and the contract start date and the contract term, given the interpretation of these effects may be questionable.

Removing interaction terms

In response to these submissions the interaction terms were removed in order to simplify the model. The removal of the interaction terms on interface type resulted in very little effect on the goodness-of-fit. Removing the interaction terms on ESA throughput had a quite modest effect, while removing of the interaction terms on contract start date and contract term had a

larger effect. However, given questions raised about the interpretation of the contract term variable and the quality or interpretation of the data for contract start date, we considered it appropriate to exclude the effects for these variables also.

When the conditioning variable interaction terms were removed, the statistical significance of the coefficient on contract term became relatively weak, with a t-statistic of less than 3. This variable became a candidate for exclusion.

Removing the conditioning variable interaction terms had a considerable influence on the coefficient values for the main effects of the outputs, log capacity and log distance. The coefficient on log capacity was reduced to about 0.8 from close to 2. A coefficient value less than 1 may be more reasonable. Further, the coefficient on log distance was reduced from around 0.3 down to about 0.07. In chapter 6 we show that these smaller coefficients are consistent with economies of scale, and the implied measures of overall economies of scale are quantified.

Structural difference in pricing 2Mbps services

One reviewer suggested that there are structural differences between the pricing of 2 Mbps links and all other DTCS services, because in the majority of cases, 2 Mbps services actually bundle together regulated tail-end services and exempt transmission services.

We could not directly test whether there is a structural difference in pricing 2Mbps services due to tail-end bundling as we do not know which 2Mbps services have bundled tail end services and which do not.

However, we have investigated the proposition about structural differences in relation to the pricing of 2Mbps services by introducing into the model an indicator variable for contracts with 2 Mbps capacity together with an interaction term between this indicator variable and the variable 'log distance'. We also tested using the indicator variable 2 Mbps services without the interaction effect. When both the main effect and the interaction term were included, both effects were statistically significant but opposite in sign. When just the main effect was included, it was insignificant. This is shown in the first model in Table D.4. These results suggest that the main effect and the interaction effect relating to 2 Mbps services were highly correlated, partly offsetting each other. We interpret the results as providing no indication that there are structural differences in pricing 2Mbps services due to tail-end bundling.

The reviewer has also suggested that 2 Mbps links are priced according to a price list at set price points. However, we were unable to undertake tests of the implications of this proposition.

Results and preferred models

Table D.4 shows the econometric results using the random effects model (estimating using ML-RE), using the full sample of contracts on exempt routes and without centering the data. In all of the models shown:

- There are no interaction terms between conditioning variables and the outputs, and

- Firm-specific fixed effects are included instead of the ‘Tier 2’ variable (and associated interaction terms).³⁶

The first model shown includes the contract term and the 2 Mbps indicator variable. Both of these variables are insignificant and the chosen level of significance. The second model shown excludes these two variables. The third model also excludes the contract start date variable. Some observations follow.

Provider fixed effects

The provider fixed-effects are almost all significant. [REDACTED] is an exception in some of the models. The joint parameter tests for the provider-specific fixed effects strongly reject the null hypothesis that the coefficients on these variables are all equal to zero. The base provider [REDACTED] is also at the median of the distribution of fixed effects. This does not appear to support the claim that market power effects are important within this sample, unless such effects are subsumed within the ESA throughput effect (as suggested by one stakeholder) or the interface type effect.

[REDACTED]. Since it was found that outliers were disproportionately represented among the smaller providers, these results tend to suggest that the most extreme of the provider-specific fixed effects are capturing part of the influence of outliers and thereby assisting to correct for differences in data quality. Aside from this type of influence, the provider-specific fixed effects may reflect differences in efficiency, product differentiation, market power, or possibly other factors.

Trend in prices

In the second model the coefficient on the contract start date is equal to 0.00005. This coefficient can be loosely interpreted as a daily rate of change in prices in percentage terms, which is equivalent to approximately 1.8 per cent per year. However, one industry stakeholder raised concerns about the quality of the data for contract start date. Given the uncertainty about the reliability of the price trend estimate provided by the *contract start date*, it may be desirable to exclude this variable, even though we note it is highly significant in Model 2. Model 3 shows the effect of excluding this variable.

Route and ESA throughput

Some questions were raised by industry stakeholders about the interpretation of the coefficients on *route throughput* and *ESA throughput*. We viewed the negative coefficient on route throughput as reflecting economies of scale in DTCS infrastructure if providers share

³⁶ The numbering system here is not the same as the codes for providers used elsewhere in this report because the providers on exempt routes are number sequentially from 1 to 9 (removing gaps where data for a provider is not present in the sample).

facilities. ESA throughput measures the total purchased wholesale transmission capacity at particular ESAs. In regard to the interpretation of this effect, we suggested in the draft report that the positive coefficient on this variable may be due to capacity constraints at exchanges in ESAs with higher density telecommunications traffic. However, we recognise the need to be cautious when making interpretations of this kind, particularly since it is likely that wholesale transmission throughput is small relative to the amount of self-supplied transmission traffic, as one stakeholder pointed out.

In light of the questions surrounding the interpretation of these two effects, and the way in which they work in opposite directions, we tested a model in which these two variables were excluded. The results are shown as Model 4 in Table D.4. Aside from the exclusion of these two variables, Model 4 is in other respects the same as Model 3.

The results indicate that the removal of these two variables does little harm to the fit of the model. Comparing Model 3 and Model 4, the R^2 is reduced slightly from 0.6819 to 0.6767, and the RMSE and MAE are increased slightly from 0.5267 and 0.3711 respectively, to 0.5316 and 0.3767. On the other hand, the BIC (which rewards parsimony) is virtually unchanged, 9355.7 in Model 3 compared to 9355.6 in Model 4. Reductions in the BIC mean a better fit and a reduction of 10 or more is considered to be a significant improvement. This result suggests there is not a material change in the goodness-of-fit resulting from the exclusion of the *route throughput* and *ESA throughput* variables.

Diagnostic tests

Table D.5 shows a set of diagnostic tests for the five models presented in Table D.4. The tests relating to the residuals indicate:

- *Normality of Residuals:* Again, the formal statistical tests reject the null hypothesis of normally distributed residuals.³⁷ The IQR tests shows that the distribution of the residuals has fatter tails than the normal distribution. The sample size in this study is considered to be sufficiently large that this result should not be an issue of concern.
- *Homoscedasticity:* Again the tests shown that the residuals are heteroscedastic.³⁸ In part this has been addressed by adopting high statistical thresholds of significance for the inclusion of variables in the model.
- *Observations with Undue Influence:* A relatively high proportion of observations have a substantial degree of influence. Among the five models shown, this varies between 3.2 per cent and 3.6 per cent.

These outcomes are similar to all of the other models tested using the 2014 dataset, and may in part relate to shortcomings in the quality of data in some instances, and may also be influenced by the lack of availability of some variables relevant to price formation. These are matters that cannot be addressed within the scope of this study.

The tests that relate to model specification include:

³⁷ Doornik-Hansen test.

³⁸ Breusch-Pagan/Cook-Weisberg test.

-
- *Multicollinearity*: The method used here to detect high multicollinearity is the number of variance inflation factors (VIFs) that exceed 10. The models shown here all have five variables with VIF scores greater than 10. These five variables are the outputs and their higher order terms. Since the outputs are always positive it is to be expected that they are correlated with their higher order terms. Thus multicollinearity does not appear to be a problem in these models.
 - *Misspecification*: The RESET test of the null hypothesis that there are no omitted variables continues to be rejected. Almost certainly there are relevant variables not available in the dataset which give rise to the omitted variable problem. The link test of the null hypothesis that the dependent variable is not misspecified is accepted in most models, and strongly so in Models 3 and 4. This result tends to support those two specifications ahead of the others.

Table D.4: Random effects models (2015 data, full sample, ML-RE estimator)

<i>Predictor</i>	Model 1 (incl. contract start date, contract term & 2 Mbps)		Model 2 (excl. contract term & 2 Mbps)		Model 3 (excl. contract start date)		Model 4 (excl. contract start date, route t'put & ESA t'put)	
	<i>Coeff.</i>	<i>t-stat.</i>	<i>Coeff.</i>	<i>t-stat.</i>	<i>Coeff.</i>	<i>t-stat.</i>	<i>Coeff.</i>	<i>t-stat.</i>
constant	5.7186	33.56	5.68261	35.07	4.78394	44.67	4.95344	57.57
log capacity	0.4888	31.84	0.49789	43.07	0.49225	42.52	0.49147	42.51
log distance	0.0968	4.07	0.09831	4.14	0.09500	3.98	0.11703	4.97
0.5(log capacity) ²	-0.0346	-10.05	-0.03596	-12.48	-0.03523	-12.19	-0.03503	-12.13
0.5(log distance) ²	0.0137	2.10	0.01319	2.03	0.01369	2.09	0.01295	1.97
(log capacity)(log distance)	-0.0040	-2.63	-0.00375	-2.47	-0.00366	-2.41	-0.00472	-3.17
log route t'put	-0.0185	-3.36	-0.01791	-3.27	-0.01966	-3.57	.	.
log ESA t'put	0.0324	4.00	0.03327	4.11	0.03027	3.72	.	.
contract start date	-0.0001	-7.35	0.00005	-7.36
contract term	0.0009	2.23
route class 2 (Metro)	0.1729	2.34	0.17166	2.33	0.17391	2.35	0.22019	2.98
route class 3 (Regional)	0.3208	5.50	0.31876	5.48	0.31498	5.38	0.32620	5.54
Provider #1	[REDACTED]							
Provider #3								
Provider #4								
Provider #5								
Provider #6								
Provider #7								
Provider #8								
Provider #9								
interface-type 3 (SDH)								
2 Mbps service	-0.0259	-0.95
$\alpha(u)$	0.3175		0.3166		0.3187		0.3247	
$\alpha(e)$	0.4272		0.4276		0.4291		0.4286	

DTCS Benchmarking Model

<i>Predictor</i>	Model 1 (incl. contract start date, contract term & 2 Mbps)		Model 2 (excl. contract term & 2 Mbps)		Model 3 (excl. contract start date)		Model 4 (excl. contract start date, route t'put & ESA t'put)	
	<i>Coeff.</i>	<i>t-stat.</i>	<i>Coeff.</i>	<i>t-stat.</i>	<i>Coeff.</i>	<i>t-stat.</i>	<i>Coeff.</i>	<i>t-stat.</i>
<i>Goodness-of-fit</i>								
R ² *	0.6825		0.6837		0.6819		0.6767	
BIC	9322.2		9310.6		9355.7		9355.6	
RMSE (based on ue)	0.5264		0.5253		0.5267		0.5316	
MAE (based on ue)	0.3701		0.3707		0.3711		0.3767	
<i>Joint significance tests</i>								
	<u><i>chisq</i></u>	<u><i>p-value</i></u>	<u><i>chisq</i></u>	<u><i>p-value</i></u>	<u><i>chisq</i></u>	<u><i>p-value</i></u>		
2 nd order output terms (df = 3)	134.3	0.0000	198.5	0.0000	189.8	0.0000	196.9	0.0000
Route classes (df = 2)	42.3	0.0000	41.9	0.0000	39.9	0.0000	36.9	0.0000
Provider fixed effects (df = 8)	1073.7	0.0000	1072.1	0.0000	1086.7	0.0000	1087.5	0.0000

Source: Economic Insights estimation results.

Notes: * Squared correlation between fitted and actual dependent.

Table D.5: Statistical tests, Random effects models (2015 data)

	Model 1 (incl. contract start date, contract term & 2 Mbps)		Model 2 (excl. contract term & 2 Mbps)		Model 3 (excl. contract start date)		Model 4 (excl. contract start date, route t'put & ESA t'put)	
	Stat.	P-value*	Stat.	P-value*	Stat.	P-value*	Stat.	P-value*
Normality of residuals								
Doornik-Hansen ⁽¹⁾	4288.6	0.0000	4012.7	0.0000	4138.0	0.0000	4084.1	0.0000
IQR (% severe outliers) ^{(2)†}	1.75%		1.79%		1.91%		1.81%	
Influential observations								
Outliers ^{(3)†}	1.80% ^a		1.83% ^a		1.77% ^a		1.80% ^a	
High leverage ^{(4)†}	1.85% ^a		1.60% ^a		1.82% ^a		2.63% ^a	
Influential observations ^{(5)†}	3.25% ^a		3.58% ^a		3.43% ^a		3.44% ^a	
Homoscedasticity								
Breusch-Pagan/Cook-Weisberg ⁽⁶⁾	1318.3 ^a	0.0000	1293.7 ^a	0.0000	1344.4 ^a	0.0000	1407.2	0.0000
Multicollinearity								
# VIF scores > 10	5/21		5/19		5/18		5/16	
Misspecification								
RESET ⁽⁷⁾	15.64 ^a	0.0000	13.46 ^a	0.0000	8.71 ^a	0.0000	9.60 ^a	0.0000
Link test ⁽⁸⁾	1.66 ^a	0.096	1.97 ^a	0.049	0.64 ^a	0.525	0.53 ^a	0.595
Joint parameter tests								
Higher-order output terms ⁽⁹⁾	134.3	0.0000	198.5	0.0000	189.8	0.0000	196.9	0.0000
Route class effects ⁽⁹⁾	42.3	0.0000	41.9	0.0000	39.9	0.0000	36.9	0.0000
Provider-specific effects ⁽⁹⁾	1073.7	0.0000	1072.1	0.0000	1086.7	0.0000	1087.5	0.0000

Note: * Null hypothesis is rejected, as a standard procedure, in these tests, if P-value is less than 0.05. Equivalently, the reported statistic exceeds the critical value for that statistic; † Percentage of $n = 6767$ observations; (1) $\chi^2(2k)$ where $k = 22$ for 1st model, and $k = 20$ for 2nd model and $k = 19$ for 3rd model. (2) Severe outliers represent about 0.0002% of a normal distribution; (3) Studentized residual > 3; (4) Hat value > $3k/n$; (5) Cook's D > $5 \times$ average Cook's D; (6) $\chi^2(1)$; (7) Via powers of the dependent variable, $F(3, n-k-3)$; (8) t -statistic on \hat{h}^2 ; (9) $F(r, n-k-r)$, where $r =$ number of parameters tested, and $r = 3$ for higher-order output terms, $r = 2$ for route classes, and $r = 8$ for provider-specific effects. (10) $\text{chibar}^2(1)$; ^a Approximate, based on OLS regression of $(y - \hat{u}_i)$ on the predictors.