Use of econometric models for cost assessment at PR19

: vivideconomics ARUP

Contents

Executive summary

This report presents recommendations on the use of econometric models in cost assessment at PR19. It extends earlier work[1] by Vivid Economics and Arup on the causes of wholesale wastewater costs and offers guidance on two connected aspects of cost assessment.

Understanding and modelling wholesale wastewater costs in bioresources and enhancement, two areas of spending not considered in detail in previous work.

Use of models in cost assessment including the employment of suites of models to predict totex for AMP7 and the setting of an efficiency challenge based on modelling evidence.

Earlier work by the project team showed how engineering insights can underpin more robust econometric models of costs and highlight their limitations. The starting point for model development was a set of engineering narratives that describe how factors drive costs and as well as the likely magnitude of their impacts. These narratives framed hypotheses that could be tested in models. In the June 2017 report, the approach yielded far more robust models of wastewater base costs than were used at PR14, accounting for drivers related to drainage, economies of scale, treatment quality and urbanisation. This report applies the approach to bioresources and enhancement, two critical areas for PR19.

The approach yields a set of clearly motivated and statistically robust, albeit still imperfect models that can be used at PR19. Models that comprise a variety of cost lines show statistically significant relationships corresponding to the main engineering narratives reviewed. Unavoidable errors and biases in the models can be moderated by the adoption of diverse suites of models, encompassing a range of explanatory variables and multiple splits of the value chain.

[1] Understanding the exogenous drivers of wholesale wastewater costs in England & Wales, Arup and Vivid Economics, 2017.

An appropriate level or form of efficiency challenge remains uncertain and will depend on using model and data quality, especially for enhancement spending. Given the paucity of data and the complexity of the service, even balanced suites of high quality models will contain error. Analysis shows that percentile challenges such as the upper quartile are not robust to different choices of equally valid models or to measurement errors in data. There is particular sensitivity to enhancement models, where model explanatory power is weakest and most uncertain, and where companies tend to outperform in areas where they spend most. To set an appropriate level and form of efficiency challenge at PR19, Ofwat can assess model and data quality to separate unexplained variation in costs into differences in relative efficiency and modelling error.

Data quality remains a concern, with key scale drivers measured with margins of error of up to 20%. Monte Carlo simulation using the data confidence grades reported by companies shows that a quarter of the difference between average and upper quartile efficiency scores could be explained by measurement error. This does not account for errors in cost lines, errors in data from outside the datashare, or inconsistencies in confidence grading between companies, all of which will exacerbate the effects of measurement error on projections of efficient costs. Attenuation bias resulting from measurement error will further affect model residuals, causing companies with 'large' explanatory factors to appear less efficient. This suggests making improvements to data quality and the consistency of data confidence grading is a priority for PR19. With a more complete understanding of measurement error, Ofwat will be able to account for it more rigorously in its choice of cost assessment models and its use of these to assess efficient costs.

For bioresources, new data highlights wide variation in land availability between companies that would be expected to drive disposal costs but surprisingly is not correlated with disposal activity. Spatial analysis reveals stark differences between companies' access to land close to production centres for the disposal of sludge, which would be expected to explain the transport work done by companies and hence their costs of disposal. However, an estimate of the expected distance to land disposal bears little relation to the distance reported in industry data and also performs poorly in models. In order to develop more robust models of bioresources costs, more clarity is required on the choices companies make over disposal routes and the way income from farmers is accounted for in the data.

Modelling is not suitable for some enhancement lines. The PR14 enhancement models produce highly unstable estimates of company unit costs, a reflection in most cases of the small samples and unrepresentative data on enhancement projects and their costs. For treatment quality improvements mandated by permit tightening, and for spending to accommodate, it is possible to improve data quality and hence models by merging enhancement lines. However, in some areas, data limitations preclude reasonable statistical estimation. Areas where modelling will not be viable include spending to meet phosphorous permits below 1mg/l, because novel and site-specific technology may be installed, and storage, where historical data on spending is dominated by a single company.

More aggregated models for bioresources and enhancement spending account for trade-offs between spending in these areas and others. Spending on certain bioresources activities can be substituted with spending on sewage treatment, while enhancement spending can in some circumstances be replaced with more intensive operation and maintenance of existing assets. As a consequence, company costs in these areas tend to be explained less well by disaggregated models than by more aggregated cost lines. It would be better to use aggregated models alongside service-specific models for these two areas, with the qualification that not all enhancement activity can be aggregated.

To set an appropriate efficiency challenge requires an assessment of model and data quality, both of which remain uncertain. Even balanced suites of high quality models contain error. Analysis shows that percentile challenges such as the upper quartile are not robust to different choices of equally valid models or to measurement errors in data. Furthermore, the approach taken to modelling means that historical outperformance will be concentrated in enhancement, where model explanatory power is weakest and most uncertain, and where companies tend to outperform in areas where they spend most. To set an appropriate level and form of efficiency challenge at PR19, Ofwat can assess model and data quality to separate unexplained variation in costs into differences in relative efficiency and modelling error and can use this information to set an appropriate level and form of efficiency challenge in PR19.

SECTION 01

Introduction

1.1 TEAM

Arup and Vivid Economics were commissioned by United Utilities to extend the analysis presented in their June 2017 report, *Understanding the exogenous drivers of wholesale wastewater costs in England & Wales*, referred to in this document as 'the June 2017 report'. The interdisciplinary team worked independently with peer review provided by Dr Ralf Martin from Imperial College and Arup subject matter experts.

1.2 OBJECTIVE AND STRUCTURE OF THIS REPORT

The objective of this work is to produce recommendations for the use of models in wholesale wastewater and the wider cost assessment at PR19, drawing on the June 2017 report and noting the methodology set out by Ofwat in December 2017. In support of the recommendations, it provides new evidence on the drivers of wholesale costs in wastewater enhancement and bioresources, which the earlier work did not explore, and considers how to synthesise evidence across candidate models.

The remainder of the report is structured as follows.

Section 2

Section 2 presents evidence on bioresources costs.

Section 3

Section 3 considers enhancement costs in AMP7.

Section 4

Section 4 combines the new models in suites.

Section 5

Section 5 investigates the extent of measurement error.

Section 6

Section 6 discusses the selection of an appropriate efficiency challenge.

SECTION 02

Bioresources

This section presents evidence on the drivers of efficient spending on bioresources and the extent to which these can be explained in econometric models.

It is structured as follows:

Section 2.1

Sets out engineering narratives on the determination of efficient bioresources costs and develops an exogenous driver of disposal and treatment costs through spatial analysis of sludge production and land bank data.

Section 2.2

Assesses econometric evidence of the significance of different drivers of costs.

Section 2.3

Concludes with recommendations for PR19.

2.1 ENGINEERING ASSESSMENT

2.1.1 NARRATIVES

The cost of providing bioresources services in a region reflect the capital and operational costs associated with treating and disposing of sewage sludge. The principal assets involved are sludge treatment centres, incinerators and tankers used in inter-site transfer ('intersiting') and disposal of sludge to land. Table 1 lists critical exogenous drivers of sludge cost and the associated engineering narratives according to which they affect efficient costs.

There is no exogenous driver that corresponds exclusively to treatment quality over the long run. In contrast to wastewater treatment, where permit tightening can drive more advanced treatment quality which in turn affects efficient costs (see the June 2017 Report), enhanced treatment quality is for the most part either an endogenous response to land availability or an unforced management decision justified by energy recovery or reduced water content. As a consequence, no exogenous driver in Table 1 affects treatment quality exclusively. However, due to its fixed nature, once advanced treatment capacity has been installed it may be the cheapest way of responding to changes in land bank.

Costs can be substituted between wastewater treatment and bioresources to a significant degree. For example, company choices over the type of thickening and dewatering process at each sewage treatment works affect the volume of sludge amount produced and its consistency conveyed between sludge treatment centres, which in turn affects bioresources costs.

Arable land is the cheapest disposal route, offering significantly greater and lower cost capacity per hectare than grassland, so its availability affects the costs of both treatment and disposal. The costs associated with disposal to arable land tend to be less than that to grassland for a variety of reasons including: the presence of livestock manures reducing pastures' capacity to receive sludge; more onerous processes for handling and application mandated by the Safe Sludge Matrix (ADAS, 2001); and, higher susceptibility to interruption during periods of rainfall. This means companies with less suitable arable land available near sludge treatment centres will either face higher costs of transporting sludge longer distances to land or find it efficient to adopt more expensive treatment and disposal routes.

DRIVER	VARIABLE	ACTIVITY	NARRATIVE
Volume produced	Sludge load produced (tds)	Intersiting, treatment, disposal	Higher volume leads to higher cost in all bioresources activities
Scale at wastewater treatment works	Proportion of load treated in bands 1-3 (%)	Intersiting, treatment	Higher percentage of load treated at small works raises unit costs as they produce low percentage dry solid sludge with lower energy content and higher contaminant sludge. This requires greater preliminary sludge treatment and/or increasing intersiting requirements.
Land bank	'Work done' or optimal distance travelled	Treatment, disposal	Reduced availability of suitable land near works raises costs of transporting sludge for disposal and/or requires more expensive advanced treatment
Geography	Proportion of sparse WwTW assets (%)	Intersiting, treatment, disposal	Treatment assets in sparse areas increase transportation costs for disposal and intersiting.
Rainfall	Annual rainfall (mm)	Disposal	Periods of high rainfall leave land unavailable for sludge spreading. This requires greater disposal distances or storage capacity or the use of alternative disposal routes.

Table 1: Many cost drivers for bioresources simultaneously affect costs of intersiting, treatment and disposal activities

Source: Arup

	AVAILABLE LAND BANK POST-RESTRICTIONS (2000-2015 AVERAGE HA)						
SEWERAGE COMPANY	**TOTAL**	**ARABLE**	**GRASSLAND**	**SHARE ARABLE**	**SHARE GRASSLAND**	**TOTAL (TDS/ HA)**	**ARABLE (TDS/HA)**
Anglian	1,303,150	1,112,550	190,600	85%	15%	0.11	0.13
Northumbrian	298,925	127,725	171,225	43%	57%	0.24	0.56
Severn Trent	674,900	332,525	342,350	49%	51%	0.33	0.67
South West	354,075	78,175	275,900	22%	78%	0.12	0.54
Southern	364,450	211,275	153,150	58%	42%	0.35	0.60
Thames	423,475	277,100	146,375	65%	35%	0.89	1.35
United Utilities	331,375	53,525	277,825	16%	84%	0.54	3.35
Welsh	724,675	70,650	654,025	10%	90%	0.09	0.88
Wessex	331,175	137,425	193,750	41%	59%	0.21	0.50
Yorkshire	494,225	306,975	187,250	62%	38%	0.29	0.47

Table 2 : Total and arable landbank availability differs considerably across the industry

Note: 2000-2015 average land bank is the average of ADAS datasets from 2000, 2004, 2010, 2015.

Source: Arup analysis of ADAS data, datashare

2.1.2 ANALYSIS

The area of land available for sludge disposal within each company region can be calculated from ADAS data. ADAS provides data on the location and type of land where sludge disposal is permitted. This was overlaid with water company boundaries to calculate the stock of suitable land available in each company's area of appointment. More details on the analytical process followed can be found in the Appendix.

Data shows the amount of sludge produced relative to the area of arable land available for disposal varies by a factor of 25 between companies, creating variation in operational challenges. Table 2 reports the availability of total arable and grassland within each company area and mass of dry solids produced by each company per hectare available. It shows that Thames Water produces eight times as much sludge per hectare of land available as Anglian Water. There is also significant variation in the composition of companies' land banks: while more than three quarters of the land available in South West Water, United Utilities and Welsh Water's regions is grassland, this falls to 15% for Anglian Water. As a consequence, one company produces 25 times as much sludge per hectare of arable land as another.

The spatial analysis of sludge production and land capacity highlights the effect of this on companies' costs to dispose of sludge. To understand the effect land bank variability has on companies' operations, one can calculate the amount of 'work' required to deliver sludge from production points to land with the requisite capacity. This work variable depends on the volumes of sludge emanating from works and the area and type of land nearby. It was calculated by a process of spatial optimisation as follows (see the Appendix for more technical details):

A map was produced showing the annual capacity of each square kilometre of arable land in England and Wales to accept sludge, based on ADAS data and accounting for variation in regulatory constraints and rotation cycles.

Major wastewater treatment works and sludge treatment centres were then plotted on this map. Company volumes of sludge produced reported in the 2016/17 datashare were allocated to wastewater treatment works in proportion to the capacity of these sites. Company volumes of sludge disposed were similarly allocated to sludge treatment centres based on the capacity of each site.

A series of optimisations then allocated sludge to available land, in order to minimise the total 'work', measured in sludge tonne kilometres, involved in transporting sludge from production or treatment sites to land. Optimisation took place at a $10km^2$ resolution.

A 'long-run' optimisation minimised the work involved in transporting sludge from wastewater treatment works to nearby land. This calculated the minimum work required for each company to dispose of the sludge it produces, treating the location of sludge treatment centres as an endogenous management decision, as it is over the long term.

A 'short-run' optimisation minimised the work involved in transporting sludge to be disposed of from sludge treatment centres to land. This calculated the minimum work for each company to dispose of sludge, treating the location of sludge treatment centres as exogenous and fixed, as it is over the short term.

For the short-run variable, these optimisations produced conflicts where companies dispose of sludge onto the same area of land. To ensure analysis was robust to this, for each company, 'best' and 'worst' cases were constructed. In the best case, the company in question could freely choose where to dispose of sludge, meaning it always prevailed in any conflict. For the latter, the company could carry out sludge disposal only after all other companies had chosen their optimal sludge allocation, meaning it always lost out in any conflict. A 'central case' equal to the mid-point between these two figures was used in regression analysis.

Optimal long- and short-run distances differ significantly between companies. This is shown in Table 3. Long-run distances are lower than short-run distances for all companies, reflecting the fact that companies' sludge volumes need to be transported a longer distance for disposal when they originate from a smaller number of points. Ranges between the high and low cases are at most 10% of the central case value and are negligible for all but three companies, reflecting modest competition for land limited to a few companies. Figure 1 provides a cartographic representation of a short-run optimisation, where colours represent areas of disposal for each company, darker shading denotes greater intensity of sludge disposal in a cell and yellow shows conflicting demands from multiple companies.

SEWERAGE COMPANY	LONG-RUN DISTANCE (KM FROM WASTEWATER TREATMENT WORKS/TDS)		SHORT-RUN DISTANCE (KM FROM SLUDGE TREATMENT CENTRES/TDS)	
	CENTRAL CASE	RANGE	CENTRAL CASE	RANGE
Anglian	14.81	4.82	42.75	2.43
Northumbrian	38.29	0.00	62.15	0.00
Southern	24.83	4.07	39.89	2.79
Severn Trent	24.98	0.00	41.48	0.03
South West	14.65	0.00	17.46	0.00
Thames	52.51	0.72	54.03	0.24
United Utilities	39.06	0.21	54.25	0.18
Welsh	N/A	N/A	39.66	4.15
Wessex	24.06	0.00	34.72	0.02
Yorkshire	24.62	0.00	32.81	0.00

Table 3: 'Long-run' and 'short-run' optimal disposal distances differ considerably across the industry

Note: 'Optimal' sludge disposal allocations to landbank from sludge treatment centres and wastewater treatment works to landbank in 2015/16 in km per tds; central cases denote the average between cases where each company is the first and last to dispose of its sludge to landbank (respectively 'best' and 'worst' cases); ranges reflect the difference between km per tds in best and worst cases; Welsh Water from long-run analysis omitted due to insufficient data on wastewater treatment works.

Source: Vivid Economics

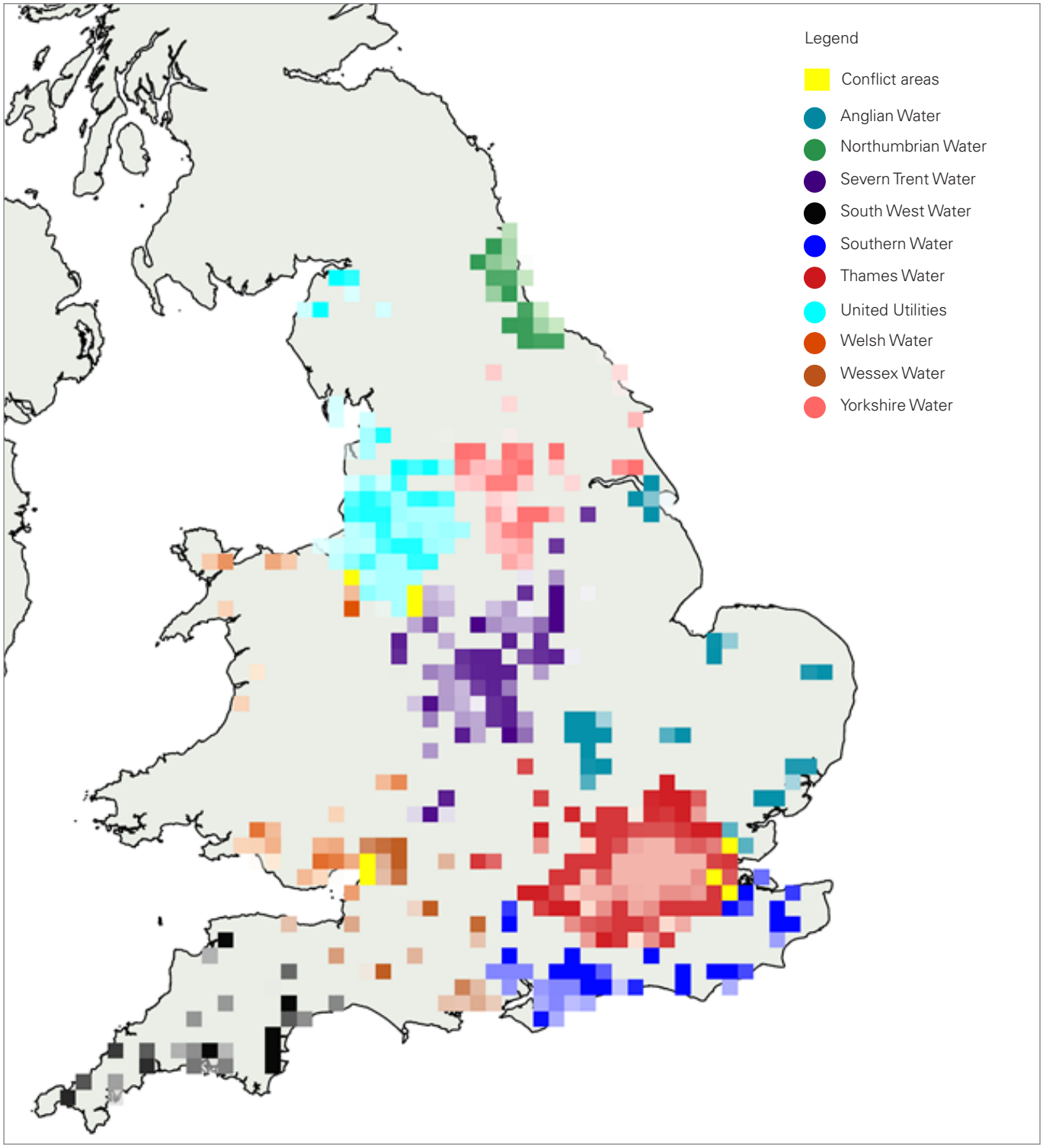


Figure 1: Estimated least cost sludge distribution results in few conflicts between companies

Note: 'Optimal' sludge disposal allocations to landbank from sludge treatment centres for the 10 companies based on sludge disposed to landbank in 2015/16; yellow areas denote areas where company allocations are in conflict with one another; darker shading represents a greater allocation of sludge to landbank.

Source: Vivid Economics

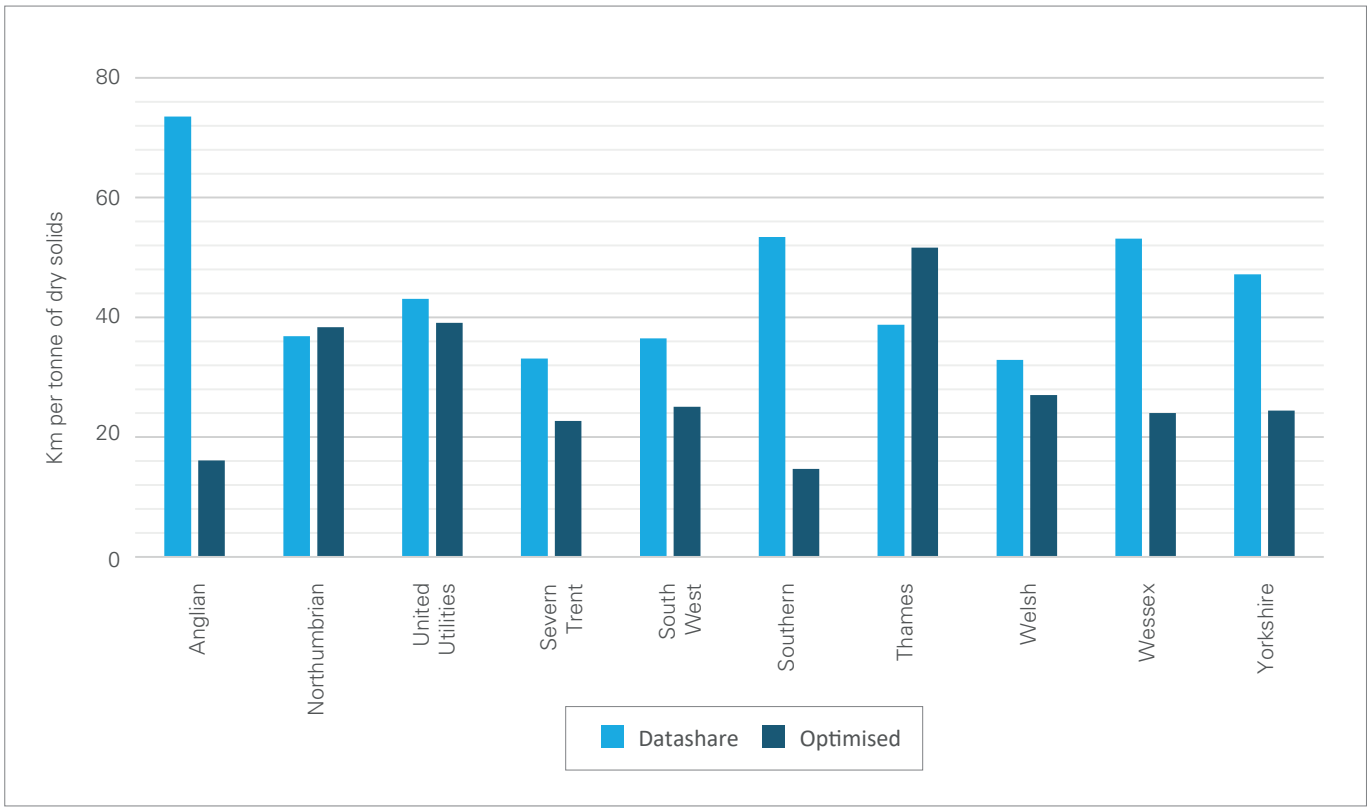


Figure 2: Optimised and actual (reported) distances from wastewater treatment works to disposal sites are not correlated across the industry

Source: Vivid Economics

Variation in optimal disposal distances is expected to drive variation in company costs worth at least £250m per AMP across the industry, or around 8% of bioresources base costs. The variation in optimal distances shown in Table 3 is expected to cause company costs to differ from what they would have been if all companies faced the same exogenous optimal distance. Using industry-wide unit costs of work done in disposal from the 2017 datashare, the value of this variation across the sector is £250m for both the short- and long-run metrics, with a majority of companies affected by more than 0.5% of wholesale wastewater totex. A more accurate estimate would inflate this figure by the difference between as-the-crow-flies distances used by the exogenous metric and actual travel distances.

Variation in optimal transportation work varies is uncorrelated with work done per tonne of sludge produced in industry data. Figure 2 highlights striking differences between long-run optimised disposal work and the actual work done reported in industry data. Not only is actual work done substantially higher than estimated work required but, when figures are normalised for differences in volume, the two metrics are uncorrelated across companies. For example, Anglian, Wessex and South West transport a tonne of sludge for disposal further, on average, than other companies, despite facing three of the lowest exogenous transport distance requirements. There is a similar lack of correlation between short-run optimised work and actual work. This may reflect a number of factors:

Efficient company choices: if companies travel further than expected in order to sell sludge to farmers or if companies who face greater landbank constraints turn to alternative disposal routes and so do relatively less transport work.

Inefficient company choices: if companies transport sludge further than is necessary.

Limitations of datashare data: industry-wide measures of work done are calculated with substantial error, as shown by an average accuracy band of 12%, and may not be calculated in a consistent manner across the industry, having only recently been reported.

Limitations in the optimised data: a key assumption of the exogenous transport work metric is that it measures disposal distances as the crow flies, so fails to account for the road network.

2.2 ECONOMETRIC ASSESSMENT

Econometric analysis tested for the presence of a statistically significant relationship between base wholesale bioresources costs and exogenous drivers of these costs. The analysis tested the short-run optimised distance variable described above alongside other variables reflecting exogenous circumstances, drawn from earlier work by this project and the 2016/17 industry datashare. Long-run optimised distance could not be used due to missing data for Welsh Water, while treatment quality metrics were not explored due to their endogeneity. Bioresources cost and driver data from previous datashares were not used because changes to service boundary definitions means older data is inconsistent. Model drivers were chosen to be consistent with the engineering account of transport, treatment and disposal costs set out in Section 4.1.

	MODEL 1	MODEL 2	MODEL 3	MODEL 4	MODEL 5
Sludge produced (I/T/D)	1.02	1.03	0.98	1.01	0.95
Optimal work done (T/D)	0.01	-0.01			
% load treated in bands 1-3 (I/T)	5.51	5.40	5.54	5.63	4.95
% sparsity (I/T/D)		0.14		0.25	
Work done in disposal / sludge disposed (T/D)			0.01	0.01	
Rainfall (D)					-0.27
Constant	-1.48	-1.40	-1.45	-2.07	0.79
R^2	0.80	0.80	0.81	0.81	0.81
VIF score	3.58	3.15	2.01	1.94	2.15
RESET test	Pass	Pass	Pass	Fail	Pass

Table 4: Intersiting and treatment variables perform well, although disposal metrics do not

Note: I denotes intersiting transport cost drivers; T denotes treatment cost drivers; D denotes disposal cost drivers. VIF scores of less than 20 are considered acceptable.

Source: Vivid Economics

Key:
- Significant at 1%
- Significant at 5%
- Significant at 10%

While intersiting and treatment variables perform well in models, the coefficients on disposal variables do not confirm expectations. Sludge produced has a positive, highly significant coefficient which is fairly stable across all five model specifications as shown in Table 4. Economies of scale coefficients are also positive, significant and stable, while sparsity is positive, but lacks significance. These conform to the narratives set out in Table 1. Short-run optimised distance measures do not perform as expected, with insignificant coefficients in models 1 and 2. This lack of correlation with costs does not refute the engineering narrative on which it is based and could simply reflect the small sample size with limited variation in most drivers over time. As explained above, it may also reflect issues within cost data, limitations of the driver, or inefficiency in company sludge disposal activity. The reported (endogenous) distance driver has positive and significant coefficients, which is unsurprising given its endogeneity. The negative coefficient on rainfall is contrary to expectations and may suggest a weak relationship between annual rainfall and temporary restrictions on landbank, with the latter driven by more acute and localised flooding events.

Bioresources models perform well against statistical criteria but have lower explanatory power than econometric models for other subservices. The model R^2 scores of around 0.8 are acceptable but lower than those witnessed in other services. This may reflect the fact that companies can substitute activities between different parts of the wastewater value chain and therefore costs between bioresources and wastewater treatment more readily than they do between other service areas, that data quality is worse, in part as a result of inconsistency between companies in cost allocation and income accounting, that a suitable exogenous land bank variable has not been identified, or perhaps that there is greater variation in efficiency for the service. Variance inflation factor (VIF) test scores indicate that multicollinearity problems are less severe in bioresources drivers than in most other base cost models, reflecting the smaller set of cost drivers. Bioresources models pass the RESET test, unlike most models in other service areas, indicating that the functional form used to include the drivers fits the data well.

2.3 RECOMMENDATIONS

Trade-offs along wholesale wastewater value chain mean aggregated models should be used as well as more granular models. While service-specific models have the potential advantage of capturing more relevant explanatory factors, aggregated approaches can allow for the fact that companies trade off activities differently between services. The importance of trade-offs between sewage treatment and bioresources means that aggregated models have superior explanatory power than granular approaches. Aggregated models should therefore be used in addition to service-specific models to estimate the cost of service provision and to understand variation in company efficiency. The importance of trade-offs between wholesale wastewater and bioresources means there is a risk that differential cost sharing incentives applied to the respective price controls will lead to inefficient company behaviours.

Service-specific models of bioresources should include exogenous drivers of treatment and inter-siting costs. Load, economies of scale and sparsity variables, which largely account for these costs, perform reliably well in models.

Significant variation in company-level drivers of treatment and disposal costs can be accounted for via an ex post adjustment to bioresources allowances. Analysis shows variation in the availability of land suitable for sludge disposal close to sludge production and treatment centres, which affects the cost of sludge disposal and trade-offs between the intensity of disposal and treatment activities. The impact of this exogenous variation on company costs is expected to exceed £250m per AMP across the industry, around 8% of bioresources botex. Though exogenous drivers of costs do not perform well in models, they could nonetheless be used as a basis for an ex post adjustment along with more granular information on unit costs of different disposal routes.

Endogenous disposal metrics should not be used in cost assessment. A general risk with the use of endogenous or asset-level metrics in cost assessment modelling is that they can reward inefficient management decisions. For bioresources disposal activities, there is a further risk of bias as companies can respond in a variety of ways to increased exogenous cost drivers, including through enhanced treatment quality, longer transportation distances or the use of incineration. The most efficient choice will depend on local factors such as opportunities to raise income from energy or fertiliser.

SECTION 03

Enhancement

This section assesses the extent to which enhancement spending in AMP7 can be explained by econometric cost assessment models at PR19.

It is structured as follows:

Section 3.1

Provides engineering evidence to support modelling, identifying drivers in industry datasets that can explain various lines of enhancement expenditure and noting where historic relationships between costs and drivers are set to change in AMP7.

Section 3.2

Considers how engineering narratives can be quantified in econometric models of costs, taking into account data limitations.

Section 3.3

Concludes with recommendations for PR19.

3.1 ENGINEERING ASSESSMENT

Some of the drivers reported in industry datasets can partially explain spending in enhancement areas, but other factors are also relevant. Table 5 lists major enhancement areas, principal drivers of costs reported in the industry datashare and other factors for which comparable data is less readily available. It shows that a large proportion of enhancement projects are specific to local assets and have multiple drivers.

Some relationships between drivers and costs observed historically are not likely to hold in AMP7. Stricter phosphorus permits of less than 1mg/l are expected to account for a substantial proportion of P-removal enhancement activity during the next price control period. Schemes of this sort often involve the adoption of new treatment technologies with costs that would vary significantly from site-to-site; the historical relationship between PE served and costs will thus no longer hold in AMP7. The Appendix provides more details on this and other areas of likely innovation in AMP7.

ENHANCEMENT AREA	COST LINES	PRINCIPAL DATASHARE DRIVERS	SELECTED OTHER DRIVERS
Growth	New development and growth STW growth	Growth in connections	System headroom
First-time sewerage	First-time sewerage	Number of properties connected by schemes; number of schemes	Sparsity, topography
Sewer flooding	Sewer flooding	Properties, combined networks	Rainfall (frequency and intensity), urbanisation, topography, asset configuration, properties at risk
Permit-led enhancement	Nitrogen (N) removal Phosphorus (P) removal at AS STW P removal at filter bed STW Ultraviolet (UV) disinfection	Population equivalent affected by permit tightening (P, BOD, UV)	Extent of permit tightening, receiving water quality, other discharges
Storage	Storage	Volume of storage, combined networks	Urbanisation, network configuration, rainfall (frequency and intensity), receiving water quality
Monitoring equipment	Flow monitoring at STWs/CSOs Event duration monitoring	Number of sites	Asset configuration

Table 5: Drivers of major enhancement cost lines

Note: AS is Activated Sludge; STW is a Sewage Treatment Works; CSO is a Combined Sewer Outflow. Other lines are conservation, chemicals pilots, groundwater, discharge relocations, odour, resilience, SEMD, freeforms. These are either not likely to be material in AMP7 or have no plausible main driver reported in the datashare. Costs associated with transferred private sewers are treated as being driven by length and other factors relevant to public sewers.

Source: Arup

3.2 ECONOMETRIC ASSESSMENT

3.2.1 DATA ASSESSMENT

Limited availability of data constrains the development of robust econometric models for enhancement spending. As noted above, the efficient level of enhancement expenditure is inherently difficult to model, with the costs of many large projects dependent on combinations of drivers that are specific to local operating conditions and legacy assets. This difficulty is severely exacerbated by some features of the data, set out below.

There is no reliable time series of observations of costs and associated drivers. Capital spending on major projects often takes place over a number of years, either in response to or anticipation of a demographic or regulatory driver that occurs at another point in time. To account for this, costs and drivers are summed over a number of years. The downside of doing this is that it reduces the number of data points and only approximately accounts for lags between spending and drivers, as not all spending accrued within the sampling period will be attributable to drivers recorded within the period and some spending accrued outside the sample will be attributable to drivers recorded within the sample. Approaches based on lagged instruments are not viable due to asset-level differences in the length of time between expenditure activity and increases in volume.

No cross-sectional information on project-level costs is reported, which, combined with a lack of time series information, reduces the sample of observations to at most ten for any cost line. As a consequence, only very simple models with a small number of explanatory factors are viable.

Spending is skewed towards a small number of companies for many enhancement lines. In at least five enhancement areas modelled at PR14, the bottom 50% of companies account for less than 10% of spending from 2011-17 (see Appendix). This means historical data may be unrepresentative of industry spending in AMP7 and it reduces the number of comparators available.

Accounting practices do not reflect the relevance of multiple drivers to projects. In many cases, enhancement projects are specified to meet a multiplicity of demands: if, for example, a company expects both population growth and permit tightening at one of its treatment works, it may be more efficient for it to carry out a single upgrade of the works in response to both drivers, rather than to upgrade the works incrementally. However, the adoption of primary driver mapping means companies often record such projects under a single line, which can bias estimates of unit costs. Similar effects may stem from companies taking different approaches to allocating costs between maintenance and enhancement, where projects involve elements of both.

3.2.2 ANALYSIS

Econometric analysis considered whether alternative specifications to those used at PR14 could produce more reliable projections of efficient costs. The PR14 enhancement models produced unstable estimates of unit costs that were influenced by a small number of companies. To understand the scope for improvement, lines of enhancement spending were assessed individually and collectively to understand first, the principal engineering drivers that determine spending, and second, the degree to which data on costs and drivers makes modelling viable. The Appendix provides more information on the performance of the PR14 models and the process followed.

Single-line models for first-time sewerage and sewer flooding may be viable. Models of single enhancement lines can be viable when activity is spread across a large number of projects undertaken by all companies in the industry, and where costs are reliably accounted for and explained by a small number of exogenous factors. Of all the enhancement lines, only first-time sewerage and sewer flooding met these conditions. For sewer flooding, the linear and loglinear functional forms used at PR14 were retained. For first-time sewerage, new linear and loglinear models that accounted for both volume of work, measured by the number of connections, and economies of scale, measured by the number of sites were tested.

Multiple-line models may be viable for treatment quality and growth. In two cases it was possible to develop potentially viable specifications by merging enhancement lines:

Treatment quality: 'Permit-led' enhancement projects often respond to multiple permit tightening events, but primary driver allocations means cost data does not reliably reflect this. To circumvent this problem with the data, loglinear and linear specifications were tested that explain combined spending on P, BOD and UV using PE affected by each type of permit tightening as independent variables.

Network and treatment growth: To some extent companies can choose between spending on network and treatment assets to accommodate growth in connections. There may also be some inconsistency in cost allocation methodologies between network and treatment across companies. At PR14, treatment growth was modelled separately while network growth was not modelled at all. By merging these two spending lines, it is therefore possible to cover a larger, more representative sample of projects undertaken by companies in response to this single exogenous driver and to remove the accounting inconsistency. Loglinear and linear models were tested that explain spending across network and treatment growth using change in connections as an independent variable.

Enhancement areas that are substitutable with base costs can be integrated with base cost models. In some areas, companies can achieve a service outcome either through spending on enhancement or through more intensive operation or maintenance of their existing assets. Where this is the case, merging relevant enhancement lines into base cost may be expected to improve the explanatory power of base cost models, especially where the base models include explanatory factors that are causally related to the enhancement lines. New models were tested that added the enhancement lines to base costs as in Table 7.

Enhancement expenditure in other areas is unsuitable for modelling, but may be amenable to special factor assessment. Expenditure in other areas, such as storage, groundwater protection and uncategorised 'freeform' lines, does not lend itself to modelling because there is insufficient data on the costs and drivers of relevant projects drawn from a representative sample. Less material lines can be accounted for in cost assessment by way of an unmodelled uplift, as was used at PR14. For substantial areas of spending, more rigorous scrutiny can be applied to evidence in company business plans, which, as suggested in Ofwat's PR19 Methodology, may be required to provide detailed information on costs, outputs, risks, optioneering and/or market testing. Such a form of scrutiny is more straightforwardly feasible for enhancement spending on specific projects, where evidence on cost benefit analysis and procurement processes can demonstrate efficiency, than it is for claims related to base costs.

DEPENDENT VARIABLE	INDEPENDENT VARIABLE(S)	R^2	GRUBBS TEST	COMMENTS
Network and treatment growth	Growth in connections	0.74 (linear) 0.82 (loglinear)	Pass	- R^2 for merged models much higher than for individual models, which was less than <0.5
Permit-led enhancement	PE affected by P, BOD, UV tightening	0.92 (linear) 0.56 (loglinear)	Pass	- Not suitable for P consents tighter than 1mg, where different technologies involved - All coefficients significant in linear model - Improvement over line-specific model may reflect accounting inconsistencies - Does not include N tightening, where sample of companies is small
Sewer flooding	Number of properties	0.5 (loglinear and linear)	Pass	- PR14 linear and log-linear specifications retained

Table 6: Area-specific enhancement model results show some scope for improving upon PR14 specifications in growth and permit-led enhancement

Source: Vivid Economics

New models were assessed for their ability to estimate efficient costs in AMP7. Where enhancement lines were integrated into base cost models, models were assessed using the same criteria applied to other base models (see Section 4.1). For new single- and multiple-line enhancement models, the assessment considered the consistency of model coefficients with engineering narratives, how well fitted costs were, using R^2, and to what extent in-sample results were sensitive to outliers, using the Grubbs test.

3.2.3 FINDINGS

Selected single and multiple-line models fit data better than PR14 models and are less sensitive to outliers. Network and treatment growth expenditure is well explained by growth in the number of connections. The permit-led enhancement model, combining schemes for P, BOD and UV, has significant coefficients for all three population equivalent drivers. The PR14 sewer flooding model performs less well, but no suitable alternatives were found.

Econometric models that combine base costs with enhancement lines perform well against statistical tests. The addition of network enhancement lines does not adversely affect the coefficients or results obtained with the base cost only network model. Odour expenditure was added to treatment base cost models without affecting model performance. Base cost model results change slightly when all network and treatment enhancement lines are included along with resilience and Security and Emergency Measures Directive (SEMD). While urbanisation is less significant, treatment economies of scale and quality variables rise in significance. These results suggest testing the inclusion of enhancement lines in base cost econometric models, and provide a compelling argument for including lines which are strongly correlated with base activity levels, for instance, resilience or SEMD. Note that the latter argument is likely to be weaker for the water services, where there is greater variation in company size.

BASE COST LINE	ENHANCEMENT COST LINES ADDED	CHANGE IN PERFORMANCE WITH ENHANCEMENT LINES
Network	Event duration monitoring equipment, CSO forward flow monitoring, private sewer spending, sewer flooding	- Number of EDM sites tested as explanatory variable: positive but not significant - Coefficients remain stable - RESET scores improve
Treatment	Odour	- Minimal change
Botex	All lines added to network and treatment models, resilience, SEMD	- Economies of scale stronger - Share of tertiary treatment more significant - Urbanisation less significant

Table 7: Expenditure lines that are substitutable with base costs can be added to base cost models without compromising model performance

Source: Vivid Economics

Data limitations make modelling unviable for other lines of cost assessment. Alternative approaches to cost assessment such as special factor or dashboard appraisal can be used in areas such as storage and P removal for permits stricter than 1mg/l, where historical cost data is either unrepresentative of industry-wide costs or simply unavailable.

3.3 RECOMMENDATIONS

Integrated modelling of base and enhancement costs lines can improve explanatory power by capturing trade-offs. These models could be further improved if industry data on maintenance and enhancement was more comparable, allowing the construction of a smoothed capex profile.

The PR14 approach to unit cost modelling can be made more robust by merging some enhancement lines. Inconsistent cost allocation practices and trade-offs between different enhancement lines can mean that drivers explain merged sets of lines better than any individual component. Areas where this can work include spending on growth and treatment quality.

Inherent difficulties in modelling enhancement expenditure mean that some lines are unsuitable for modelling and models that are used will be less robust than those for base costs. Difficulties stem from a lack of data on costs and drivers and the nature of enhancement activity in many areas, which covers a small sample of projects involving specific assets. Section 6 explains the implications of using less robust enhancement models for setting an efficiency challenge.

SECTION 04

New cost assessment models

This section presents models of wastewater costs that could be adopted for benchmarking at PR19.

It synthesises recommendations from the June 2017 report on the modelling of wholesale wastewater costs and from the assessments of bioresources costs and enhancement spending presented in Sections 2 and 3 respectively.

The remainder of the section is structured as follows:

Section 4.1

Presents individual and collective criteria for model assessment. It argues that by testing suitable models in diverse suites that fit costs along the value chain in different ways, it is possible to mitigate the models' unavoidable individual weaknesses.

Section 4.2

Introduces suites whose consistency with engineering narratives and performance against statistical criteria make them suitable for use in cost assessment.

Section 4.3

Summarises recommendations for cost assessment in PR19 based on this analysis.

A separate Appendix considers statistical approaches to optimal model weighting or triangulation.

4.1 ASSESSMENT CRITERIA

The objective for cost assessment models at PR19 is to predict the efficient costs of each company for AMP7. The purpose of cost assessment models is not to predict costs for the whole sector, but rather to protect customers and investors by allowing each company to recover its own efficient costs. This means that models should seek to produce unbiased projections of costs. To reduce the risks to all stakeholders, it is also desirable for models to reduce the expected magnitude of errors around projections.

Individual models are assessed against engineering and statistical criteria for predictive power. These are described fully in the June 2017 report. Engineering criteria consider whether model explanatory variables represent factors that will cause costs in AMP7 and whether the sign and magnitude of model coefficients are consistent with these causal narratives. These criteria thus consider models' predictive plausibility directly. Statistical criteria are more limited because they appraise models' predictive power only through models' performance in *historical* datasets. The criteria used are the statistical significance of coefficients, model stability measured using Variance Inflation Factors (VIFs) and performance against the RESET misspecification test.

No individual model will perform perfectly against these criteria, so model suites can be assessed in suites. With a large number of causal narratives to account for and limited data available, all models will predict costs with error and biases that affect companies in different ways. By choosing suites of models with different underlying assumptions or drawbacks, errors and biases can be reduced though not eliminated, which will improve the accuracy of predictions and reduce risks. The use of a diverse set of models is more likely to achieve this than a set of very similar models, whose errors and biases will be highly correlated with each other.

Diversity can be attained through the use of a variety of independent and dependent variables. For the modelling of a particular cost line, a more diverse set of models might span a wider range of independent variables, covering more engineering narratives than any single model can accommodate, or it might cover narratives in multiple ways where there is more than one suitable variable available. The use of a mix of 'aggregated' and 'disaggregated' modelling that assess costs at different levels of granularity might further add to diversity. As Figure 3 explains, more or less aggregated approaches can each have distinct advantages, so can complement each other in modelling suites.

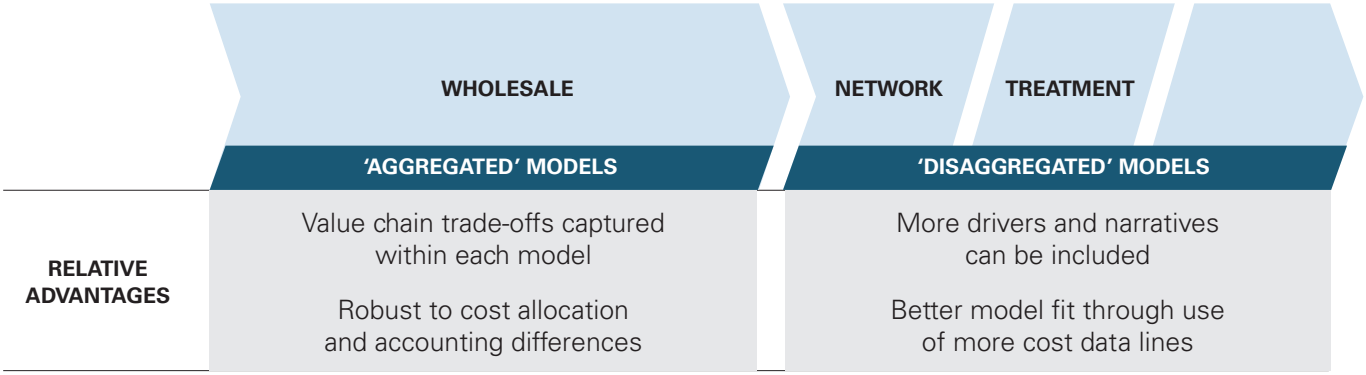


Figure 3: Aggregated and disaggregated approaches have different relative strengths and so can complement each other in modelling suites

Note: Value chain splits indicative.

Source: Vivid Economics

Figure 4: Suite selection process involves individual and collective assessment of models over multiple stages

Note: While splits containing weak models which did not contribute to diversity were dropped, at least one split included bioresources models, due to the importance of a bioresources model in setting separate price controls at PR19

Source: Vivid Economics

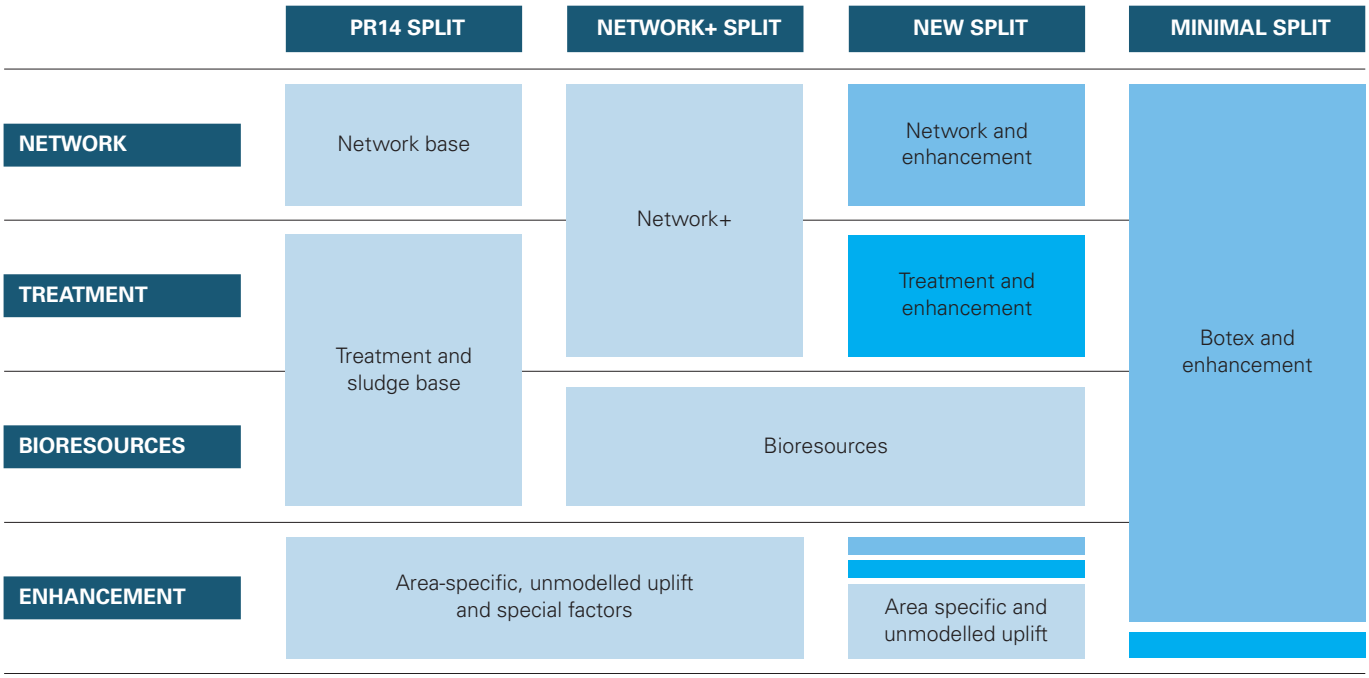


Figure 5: The four value chain splits represent a wide range of approaches from granular, bottom-up models to top-down models

Note: Within a column, costs that are shaded in the same colour are modelled together. For instance, in the 'New split', treatment base costs and enhancement lines associated with treatment are included in the same econometric model.

Source: Vivid Economics

At PR14, modelling suite diversity was limited, with a small number of models and a similar set of drivers throughout. For each service, the same set of drivers were used throughout, with botex models differentiated from treatment and sludge (T&S) models only by the inclusion of an economies of scale variable. Diversity in dependent variables was partially explored through the use of botex (top-down), and network and T&S (bottom-up) models. The principal source of diversity in the PR14 models, the use of a mix of ordinary least squares (OLS) and panel data (GLS) models, did not succeed in improving accuracy or reducing risks given the GLS models' lack of robustness.

Formal statistical approaches to assess diversity are available but not recommended to derive a suite of acceptable models. Statistical approaches treat the problem of improving predictive accuracy through multiple weighted models as analogous to that of minimising risk in a portfolio while achieving a target rate of return. The main drawback of these approaches is that they are backward looking, considering only the ability of models to estimate past costs. This means they place unduly high weight on overfitted models, which makes them poorly suited to screen acceptable models. With the same caveats, they may be informative in the weighting or 'triangulation' of models. This is discussed in more detail in the Appendix.

The selection process shown in Figure 4 assesses diversity qualitatively.

4.2 NEW MODELS AND RESULTS

The new models are organised into several 'totex splits' that explain different groupings of value-chain cost lines. The totex splits are shown in Figure 5. This use of multiple totex splits is the principal way in which a diversity of modelling approaches was achieved.

Two splits retain the PR14 enhancement modelling structure of unit cost models. The split labelled 'PR14' combines models that cover the same value chain elements as those the disaggregated model PR14. 'Network plus' separates base costs into the two PR19 price controls, network plus and bioresources.

By contrast, the 'new' and 'minimal' totex splits, contain econometric models which combine base and enhancement cost elements. As set out in Section 3, the merging of some enhancement lines into econometric models of base costs may be preferable to modelling these lines separately. The shading in the two righthand columns of Figure 5 illustrates how some enhancement costs are modelled together with base costs in these suites.

Multiple models were used for only a couple of the subservices represented in the totex splits. In some cases, engineering and statistical criteria clearly supported one specification ahead of others. This does not mean that the model was perfect for the subservice in question. For example, only one network+ model was chosen. This featured urbanisation and economies of scale drivers that were preferred to other variables corresponding to the same narratives, but did not include drivers corresponding to other narratives discussed in the June 2017 report, such as drainage and sparsity, which were not significant due to collinearity with other drivers. As a consequence, no other network+ models were included in the suite despite the one chosen model's drawbacks. By contrast, two treatment and sludge base models were chosen because both share of tertiary treatment and share of advanced tertiary treatment were equally strong candidate quality drivers, and both perform well in models. Table 8 sets out the new econometric models, based on the cost categories shown in Figure 5.

Individual models perform well against engineering and statistical criteria, although concerns around omitted variables remain. The drivers much more clearly reflect engineering arguments than the drivers used in the PR14 models. The coefficients are almost always statistically significant, and coefficient signs and magnitudes are consistent with engineering narratives. Other statistical tests results show that new models have a high degree of fit, as in PR14, but in contrast to overfitted PR14 models, the new models' variables appear to be much less collinear, based on VIF scores. This suggests that the new models are likely to be more robust than the PR14 models. Most models fail the RESET test, though more marginally than the PR14 models did, indicating continued misspecification that may require improved data collection to remedy. As explained in the June 2017 report, the use of alternative functional forms that lack an engineering basis but which fit the data and thus may pass the RESET test is unlikely to improve models' predictive power. For this reason, the models that fail the RESET test but that perform well in other respects are considered acceptable. The Appendix provides more detailed results.

4.3 RECOMMENDATIONS

The improved models proposed in this work, though still imperfect, could be used in PR19. Unlike the PR14 models, they are motivated by engineering accounts of the cost of service provision along the value chain and reflect narratives around treatment quality, drainage and urbanisation. Unlike the PR14 models, they estimate statistically significant relationships within stable specifications. However, in common with all models of the wholesale wastewater service, they remain imperfect, with some evidence of misspecification. Models also remain subject to errors in data. The effect of measurement error is discussed in more detail in Section 5.

The adoption of a diverse suite of models can improve predictive power. Due to the limited number of observations and drivers, all models are subject to errors and bias, but the use of a more diverse set of models mitigates this. Diversity can be assessed qualitatively, considering (i) the number of narratives covered by drivers, and (ii) by the mix of disaggregated and aggregated approaches to modelling the costs along the value chain. Statistical approaches to triangulation are not recommended for model screening, as they ignore engineering narratives and place undue weight on over-fitted models, but may inform model triangulation for final cost assessment.

MODEL NAME	RESPONSE VARIABLE	NO. OF MODELS USED	EXPLANATORY FACTORS (FIRST MODEL WHEN MULTIPLE MODELS USED)	DIFFERENCE BETWEEN MODELS (WHEN MULTIPLE MODELS USED)
Network base	Network base cost	1	Total sewer length, annual runoff, time fixed effects, share of urban population, constant	
Network and enhancement	Network base cost + network enhancements	1	Total sewer length, annual runoff, time fixed effects, share of urban population, constant	
Network+	Network and treatment base costs	1	Load, share of treatment bands 1-3, time fixed effects, share of urban population, share of tertiary treatment, constant	
Treatment and enhancement	Treatment base costs + treatment enhancements	1	Load, share of treatment bands 1-3, time fixed effects, share of urban population, share of tertiary treatment, constant	
Treatment and sludge base	Treatment and bioresources base costs	2	Load, share of treatment bands 1-3, time fixed effects, share of urban population, share of tertiary treatment, share of sparsity, constant	Replace share of tertiary treatment with share of advanced tertiary treatment
Bioresources	Bioresources base costs	2	Sludge produced, share of treatment bands 1-3, constant	Add share of sparsity
Botex and enhancement	Wholesale base cost + all enhancement lines*	1	Load, share of treatment bands 1-3, time fixed effects, share of urban population, share of tertiary treatment, constant	

Table 8: The suites includes a mixture of top-down and bottom-up models and cover almost all key engineering narratives

Note: * Network enhancements refers to: event duration monitoring equipment, CSO forward flow monitoring, private sewer spending, sewer flooding. Treatment enhancements refer to: odour. Base enhancement lines refer to: all network and treatment enhancement lines, resilience, SEMD. Underlined explanatory factors are logged in econometric models; total sewer length is the sum of public and private sewer lengths.

Source: Vivid Economics

SECTION 05

Impact of measurement error on cost assessment

This section considers the effect of measurement error on model results and the estimation of company efficiency scores.

It is structured as follows:

Section 5.1

Describes the scale of measurement error in industry data on costs and explanatory variables.

Section 5.2

Examines the effect of measurement error on model results by through a Monte-Carlo simulation of errors in explanatory variables.

Section 5.3

Draws out recommendations for PR19.

5.1 SCALE OF MEASUREMENT ERROR

Measurement error is an important consideration in cost assessment modelling that has not been explored in detail, either in the June 2017 report or wider industry literature. Margins of error around cost and driver data affect both the econometric modelling and efficiency analysis conducted in cost benchmarking exercises.

Measurement error in drivers can be very large and appears not to be accounted for consistently in industry data. As Figure 6 highlights, companies report wide uncertainty ranges around key scale variables, with confidence intervals of up to ±25% for total sewer length, ±9% for load and ±40% for density (not used in new models). For composite variables such as density, uncertainty is expected to be greater under the presence of measurement error in both numerator and denominator. Variation in confidence interval widths across the industry suggests that companies may report confidence grades inconsistently, despite Ofwat guidance. For example, Wessex Water has a much narrower confidence interval width around load than other companies, while United Utilities' confidence interval for sewer length is significantly wider than other companies at ±25%. This suggests that measurement errors reported in industry data are themselves subject to significant measurement error, which can compound its effects.

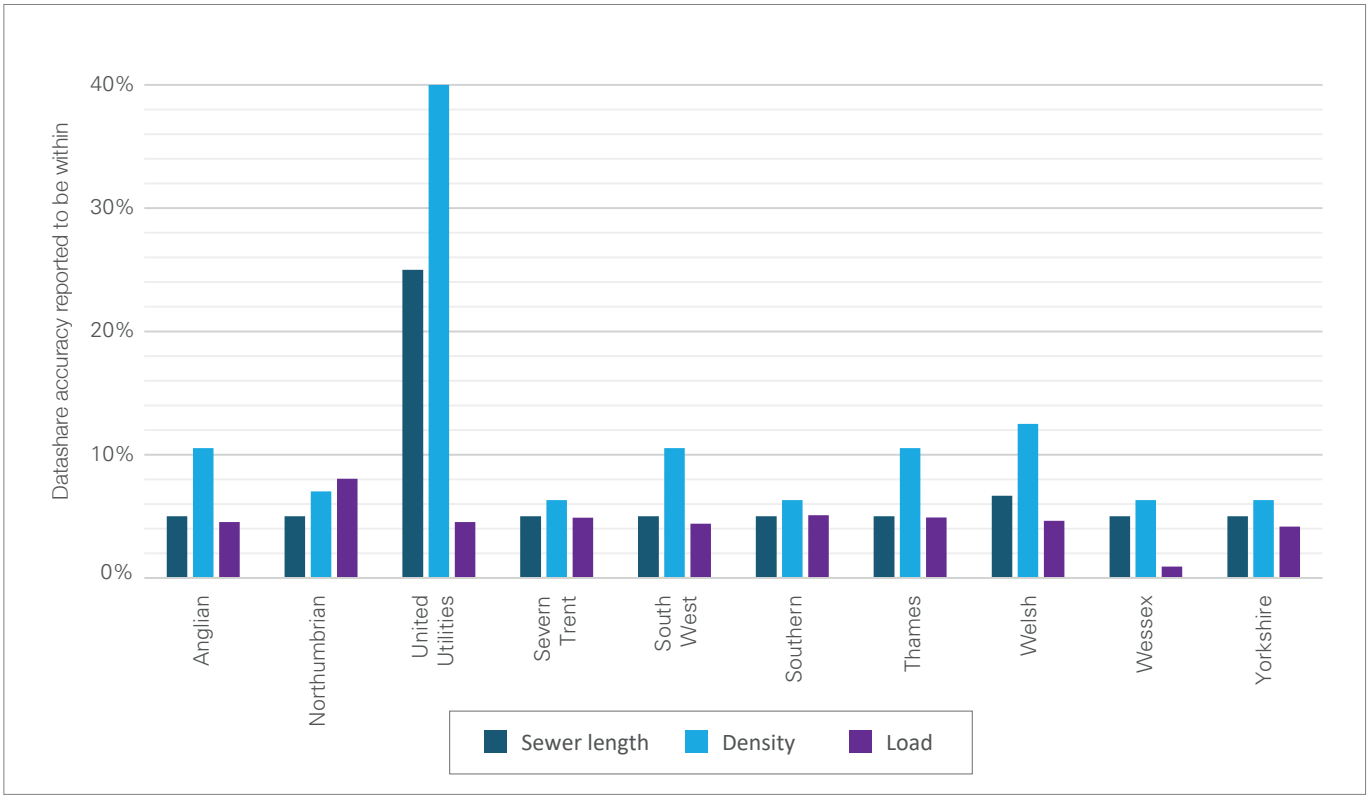


Figure 6: Important model drivers are measured with considerable uncertainty across the industry

Source: Vivid Economics analysis of 2017 industry datashare

Errors and inconsistencies in cost allocation also appear to be substantial and can cause disaggregated models to perform poorly. The large number of changes in costs lines seen in iterations of the 2016/17 datashare suggests inconsistent practice between companies in allocating costs to different value chain activities. As an indication of the significance of this, in one example 26% of costs were remapped from base to enhancement. Given the industry's limited ability to scrutinise individual company accounting practices, it is very unlikely that all such errors or inconsistencies have been eliminated in the dataset used in this report.

The effect on model results of measurement error is compounded by the persistent nature of many errors. Model variables have 'stock' characteristics if they represent quantities that change over time due to inflows or outflows. If a stock variable is measured inaccurately in the first datashare year, it is highly likely that the same mismeasurement will be made in all future years. Many key variables used in modelling, such as sewer length and resident population have stock characteristics. Even for flow variables such as load, it is likely that the assumptions or techniques that lead to errors are retained from one year to the next, leading to errors that are persistent over time. Persistent errors can have a stronger effect on model results than random errors that vary more over time.

5.2 MONTE-CARLO SIMULATION

5.2.1 ANALYSIS

Monte Carlo simulation techniques can assess the effect of uncertainty in drivers on the robustness of model coefficient estimates and efficiency scores. This technique first defines a distribution around drivers based on reported confidence intervals, before taking random draws from a uniform distribution with a width based on each company's confidence interval. These draws are then used to calculate model variables, after which the model specification is run and post-estimation statistics such as efficiency scores are calculated. The approach followed here treated errors as 'one-off' in the sense that they shock a company's driver data by the same amount each year, consistent with the idea of drivers having stock characteristics. Figure 7 shows the steps in the measurement error diagnostic procedure.

5.2.2 FINDINGS

Monte Carlo simulation results find measurement error to have a substantial effect on model coefficients. As Table 9 shows, though the signs of engineering driver coefficients are robust to measurement errors, their magnitude can vary substantially, with network length coefficients varying by more than 50% within 95% confidence intervals generated by the simulation. In general, bioresources and network model results are more sensitive to measurement error than those of botex (including some enhancement) and treatment and sludge models. As a consequence, individual company fitted costs and efficiency scores are also found to be sensitive to errors.

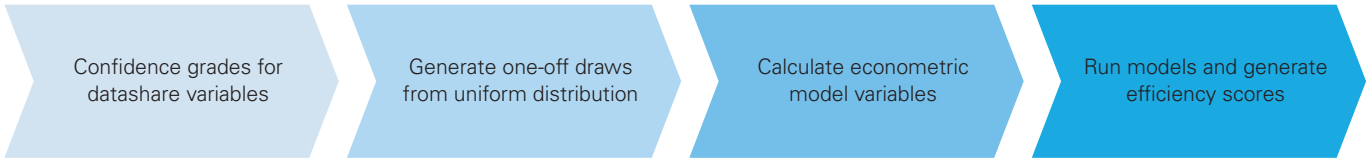


Figure 7: Monte Carlo simulation procedure for estimating the effects of measurement error on model outputs

Source: Vivid Economics analysis of 2017 industry datashare

	PR14 SPLIT NETWORK			PR14 SPLIT T&S			NETWORK+ SPLIT BIORESOURCES			MINIMAL SPLIT BOTEX+		
	Min	No ME	Max	Min	No ME	Max	Min	No ME	Max	Min	No ME	Max
Total length	0.18	0.38	0.58									
Total load				0.92	0.94	0.97				0.84	0.87	0.90
Sludge produced							0.87	1.02	1.09			
% load bands 1-3				8.40	9.59	10.87	2.19	5.42	7.71	4.66	5.80	6.91
Annual runoff	0.07	0.31	0.56									
2006/07 dummy				0.00	0.00	0.00						
2007/08 dummy				0.06	0.06	0.06						
2008/09 dummy				0.11	0.11	0.11						
2009/10 dummy				0.15	0.15	0.16						
2010/11 dummy				0.12	0.12	0.12						
2011/12 dummy	0.00	0.00	0.00	0.10	0.10	0.11				0.00	0.00	0.00
2012/13 dummy	-0.25	-0.10	0.04	0.11	0.11	0.11				0.04	0.04	0.04
2013/14 dummy	-0.12	-0.03	0.06	0.08	0.08	0.09				0.05	0.05	0.05
2014/15 dummy	-0.09	-0.03	0.04	0.09	0.10	0.10				0.07	0.07	0.07
2015/16 dummy	-0.16	-0.06	0.04	0.10	0.11	0.12				0.00	0.01	0.01
2016/17 dummy	-0.03	0.01	0.06	0.13	0.14	0.15				0.05	0.05	0.05
% urban population	0.52	0.74	0.93	1.77	2.06	2.35				0.73	1.01	1.26
% sparse				0.45	0.54	0.63						
% tertiary treatment				0.08	0.15	0.21				0.25	0.30	0.36
Constant	-2.96	-2.35	-1.65	-9.62	-9.17	-8.72	-1.80	-1.38	-0.56	-7.16	-6.69	-6.18
N	60	60	60	110	110	110	60	60	60	60	60	60
R^2	0.841	0.857	0.881	0.955	0.960	0.964	0.713	0.796	0.827	0.944	0.949	0.954

Table 9: Effect of measurement error on model coefficients is substantial in the bioresources subservice

Note: Min denotes lowest point in 95% confidence interval for coefficient generated from Monte Carlo draws; max denotes highest point in the same confidence interval.

Source: Vivid Economics

Measurement error produces ranges in assessed company costs worth many hundreds of millions of pounds, which will be compounded by further error in the efficiency challenge. Figure 8 shows the sensitivity of company fitted costs to measurement error in the 'PR14 split' suite of models. Across the industry, this is equivalent to over £520m per AMP within a 95% confidence interval in the historic data, but the effect on AMP7 fitted costs is expected be greater as many explanatory variables trend upwards over time. This range of uncertainty will be compounded by the impact of measurement error on the efficiency challenge if this is based on model efficiency scores (see Section 6.2). Variation in the upper quartile efficiency scores within a 95% confidence interval is equivalent in value to £590m if such a challenge was to be used to benchmark efficient costs.

The true effect of measurement error on allowances set through cost assessment is likely to be even greater and shared unevenly between companies. The analysis above provides only a partial account of the effect of measurement error on fitted costs and efficiency scores. Other aspects not formally modelled are likely to increase its impact on cost assessment outcomes:

Measurement error in non-datashare variables is not simulated, but likely to be material.

Inconsistencies in cost allocation, not accounted for in the Monte-Carlo analysis, affect model coefficients and efficiency scores. This is particularly important where costs are allocated between value chain elements for which cost models' goodness of fit varies, such as bioresources and network+ or enhancement and base costs.

Measurement error in explanatory variables can lead to attenuation bias, where model coefficients are biased towards zero. For the new models presented in Section 4, where all coefficients are positive, this is likely to result in 'large' companies with high driver values having understated fitted costs and hence unrealistically weak efficiency scores, while 'small' companies are conversely assigned unrealistically high scores. This contradicts a critical objective of cost assessment, which is to estimate the efficient costs of each company rather than the industry as a whole.

A full understanding of measurement error could improve model selection. Measurement error affects models' predictive power and is thus a factor to consider when choosing drivers, models and suites of models, as discussed in Section 4.1. Subject to the other criteria set out in the wastewater report, selecting drivers with low levels of measurement error will improve models, while the widespread presence of measurement error in drivers increases the value of a diverse suite containing a range of different drivers. The balance between measurement error in costs and measurement error in drivers will also affect the choice between aggregated and disaggregated models: aggregated models with a single scale variable will perform less well with measurement errors in drivers, while disaggregated models' performance declines with errors in cost allocation. Without a comprehensive view of the extent of measurement error, an assessment of this kind cannot be performed.

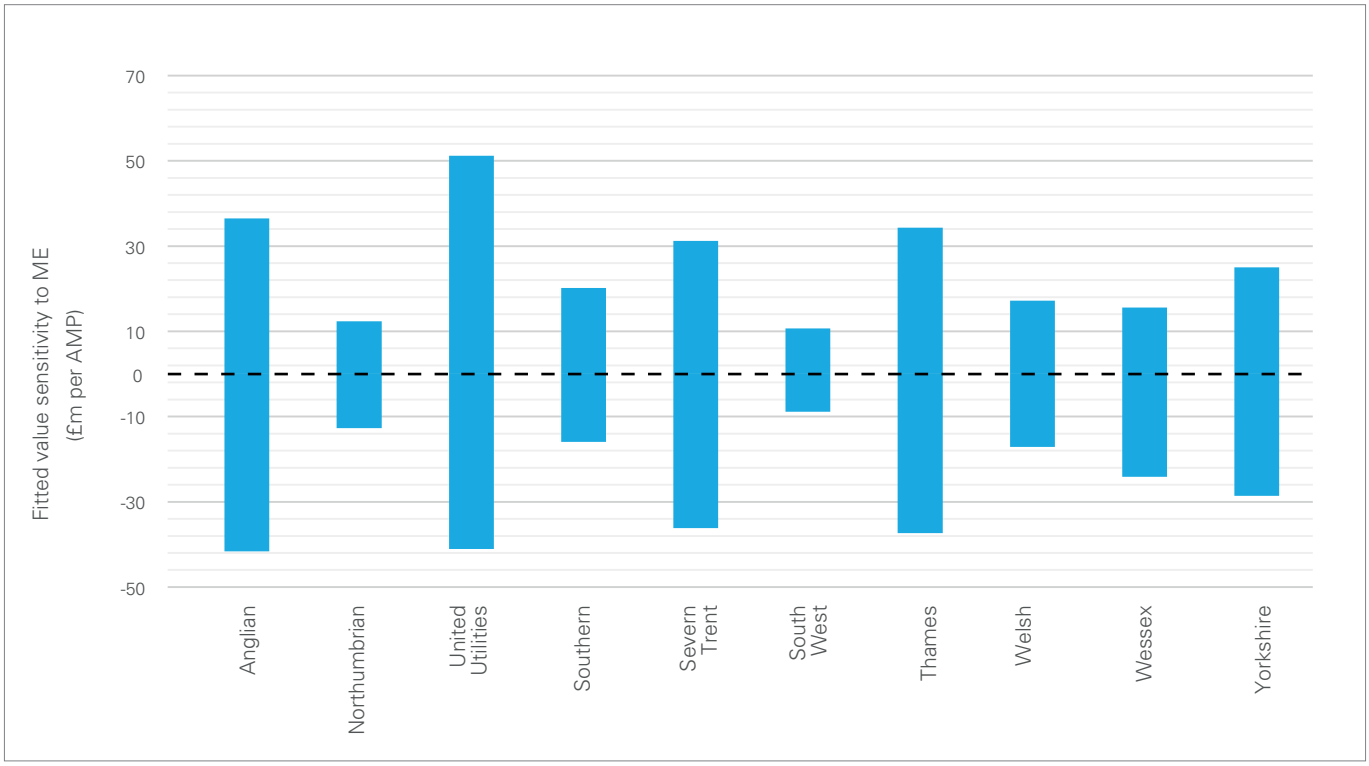


Figure 8: Monte Carlo simulation shows measurement error has a significant impact on company allowances

Note: Calculated through company 'fitted allowances' resulting from ME Monte Carlo simulation, taking the differences between these allowances over a five year period within the dataset and the 'base case' allowance generated from models without measurement error. Bars reflect 95% confidence intervals around 'base case' allowances.

Source: Vivid Economics

5.3 RECOMMENDATIONS

The effect of measurement error should be accounted for explicitly in model selection and the efficiency challenge. Both the quality of models and the interpretation of efficiency scores are affected by measurement error. With a full understanding of measurement error, Monte-Carlo analysis similar to that shown above could be used to understand the robustness of individual models, model suites, and efficiency scores.

Improvement to cost and driver data is a priority to improve cost assessment, both in the short and long term. Measurement error substantially affects both model coefficients and the efficiency scores that may be used to set the efficiency challenge. Simulation shows variation in the upper quartile caused by errors in a subset of drivers alone could shift fitted company costs by £520m per AMP and the upper quartile efficiency challenge by a further £590m. Improvement to cost data, which is not formally modelled, is likely to be particularly valuable and could be achieved at low cost in the short term.

More consistency in reporting measurement error is a prerequisite for understanding its effect on cost assessment. Confidence intervals reported vary by an implausible amount between companies: this 'error in measurement error' amplifies the impacts shown in this report.

SECTION 06

Selection of an efficiency benchmark

This section considers how modelling evidence can be used to set the efficiency challenge at PR19.

It is structured as follows:

Section 6.1

Explains the role of econometric evidence in setting an efficiency challenge.

Section 6.2

Considers evidence on how unexplained variation between companies can be decomposed between efficiency differences and model noise or bias to set a static challenge.

Section 6.3

Reviews evidence on the interpretation of time trends used to set a dynamic challenge.

Section 6.4

Concludes with recommendations for PR19.

6.1 ROLE OF MODELLING EVIDENCE IN AN EFFICIENCY CHALLENGE

The purpose of cost assessment at PR19 is to allow only efficient costs to be recovered from customers, which will require the use of an efficiency challenge. Ofwat will use econometric models to estimate a benchmark of efficient costs. To achieve this, it will apply an efficiency challenge to cost allowances produced by models, adjusting company costs downwards from their business as usual level to those of an efficient company.

This challenge will have static and dynamic components. The static component is used to adjust cost thresholds down to the frontier level observed in the data for the first year of a price control. The dynamic component then ensures thresholds remain at the frontier as companies become more productive over time, due to learning and technological progress.

Unexplained variation highlighted by econometric model results can be used to inform static efficiency challenge, but not without further evidence on the quality of the models. Econometric models show *fitted* costs explained by exogenous drivers, which can be compared to actual observed costs. This *unexplained variation* between companies' costs is a mixture of companies' relative efficiency on one hand and company-specific noise and model biases on the other. As was explained in Section 4, noise and biases are unavoidable and material features of any suite of wholesale wastewater cost models. To decompose unexplained variation into efficiency differences and other factors, two sources of evidence can be used:

Evidence on reasonable ranges of efficiency variation between companies based on past experience or evidence from other sectors;

Evidence on the contribution of driver data and modelling error to unexplained variation based on the explanatory power of models and the accuracy of underlying data.

Similarly, trends in costs can inform the dynamic challenge alongside an understanding in the source of these in models. To judge whether historical trends are likely to persist in the future, it is necessary to consider the extent to which these are explained by changes in service quality and real price effects not accounted for in the threshold, as well as gains in efficiency.

This means that suitability for PR19 of the efficiency challenges applied by Ofwat at PR14 depends on the models and underlying data that are to be used. At PR14, Ofwat used the upper quartile (UQ) of company efficiency scores as its static efficiency challenge, thus effectively attributing any unexplained variation beyond this level to noise and bias. For its dynamic challenge, it projected positive time trends in costs forward, equivalent to assuming that omitted factors, real price effects and efficiency trends would continue to evolve in AMP6 as they had in the historical database. It may transpire that these are tenable approaches for PR19, but this will depend on the models and data used: as explained below, the PR14 challenge was sensitive to the use of overfitted base cost models and enhancement models that tended to overstate outperformance.

6.2 STATIC CHALLENGE

This section confirms the importance of model and data diagnostics in setting a static efficiency challenge. It tests the sensitivity of the PR14 upper quartile challenge to changes in models and driver data, finding considerable variation in the level of the challenge despite no underlying changes in efficiency in the data. This underlines the importance of setting a challenge only on the basis of an understanding of model predictive power and data accuracy. Stochastic Frontier Analysis (SFA), an alternative approach to understanding relative efficiency using models, lacks robustness in this dataset due to serial correlation between explanatory variables: the Appendix provides more discussion of this.

Equally reasonable modelling suites can produce different levels of unexplained variation, including at the upper quartile. Table 10 shows efficiency scores generated by the new models. As Section 4.2 explains, models in the suites are deliberately chosen to balance errors and biases and the suites themselves are considered equally valid, but they nonetheless produce markedly different distributions of unexplained variation. The position of the frontier company varies by five percentage points and, while the upper quartile[2] score is relatively stable, this is coincidental as efficiency scores generated from base cost assessment only vary by more than ten percentage points. This variation is observed despite there being no underlying difference in relative efficiency in the data. Thus it is uncertain whether a percentile challenge such as the upper quartile will identify the frontier.

Most of the overall outperformance in historical data stems from the approach taken to enhancement modelling. The industry-wide PR14 efficiency challenge would have been just £840m, rather than £2,000m, if enhancement outperformance levels were equal to those in base cost models. As shown in Table 10 the upper quartile challenge generated by base cost models alone is substantially less than that overall: indeed, for three of the four new suites, base cost models give a negative challenge at the upper quartile . Outperformance in enhancement is to a large extent a product of the approach adopted in enhancement models. As noted in Section 3, disaggregated enhancement models often cover individual activities dominated by one or two companies: due to economies of scale and possibly self-selection towards areas of comparative advantage, these companies tend to outperform the models in these areas. This outperformance can be particularly pronounced when the models have relatively weak explanatory power. As a consequence, eight out of the industry's ten companies outperformed on enhancement at PR14.

Greater use of model diagnostics could have informed a more judicious choice of efficiency challenge at PR14. Diagnostics of the PR14 models presented in the June 2017 Report reveals that the base cost models were overfitted, which tends to reduce the efficiency challenge, but the approach to enhancement models will suggest high levels of potential outperformance. This made the value of the upper quartile challenge unstable and raises questions over the legitimacy of applying the challenge equally to companies with differing mixes of base and enhancement costs. Tellingly, the PR14 wastewater efficiency challenge was 4 percentage points larger than that in water, with most of the difference attributable to enhancement spending. As 10 of the 18 companies involved in the water sector are also involved in providing wastewater services, it would be surprising that the relative performance of the frontier company was so radically different.

[2] This is a possible result where the distribution of efficiency scores is skew. For this reason it is also potentially a valid efficiency challenge.

	PR14 WATER	PR14 WASTE	NEW: PR14 SPLIT	NEW: NETWORK PLUS SPLIT	NEW: NEW SPLIT	NEW: MINIMAL SPLIT
Frontier			0.784	0.817	0.761	0.791
Full UQ score	0.935	0.896	0.869	0.871	0.879	0.870
Base cost only UQ score	0.941	0.956	0.918	1.014	1.030	1.021

Table 10: Efficiency scores across PR14 and new model suites presented in Section 4, based on full and base cost only methodologies

Note: Full UQ scores based on PR14 Upper Quartile efficiency score methodology; Base cost only UQ scores use the same methodology, but omit enhancement expenditure and allowances in the calculation of efficiency scores

Source: PR14 models: Ofwat wholesale UQ efficiency challenge files.
New models: Vivid Economics

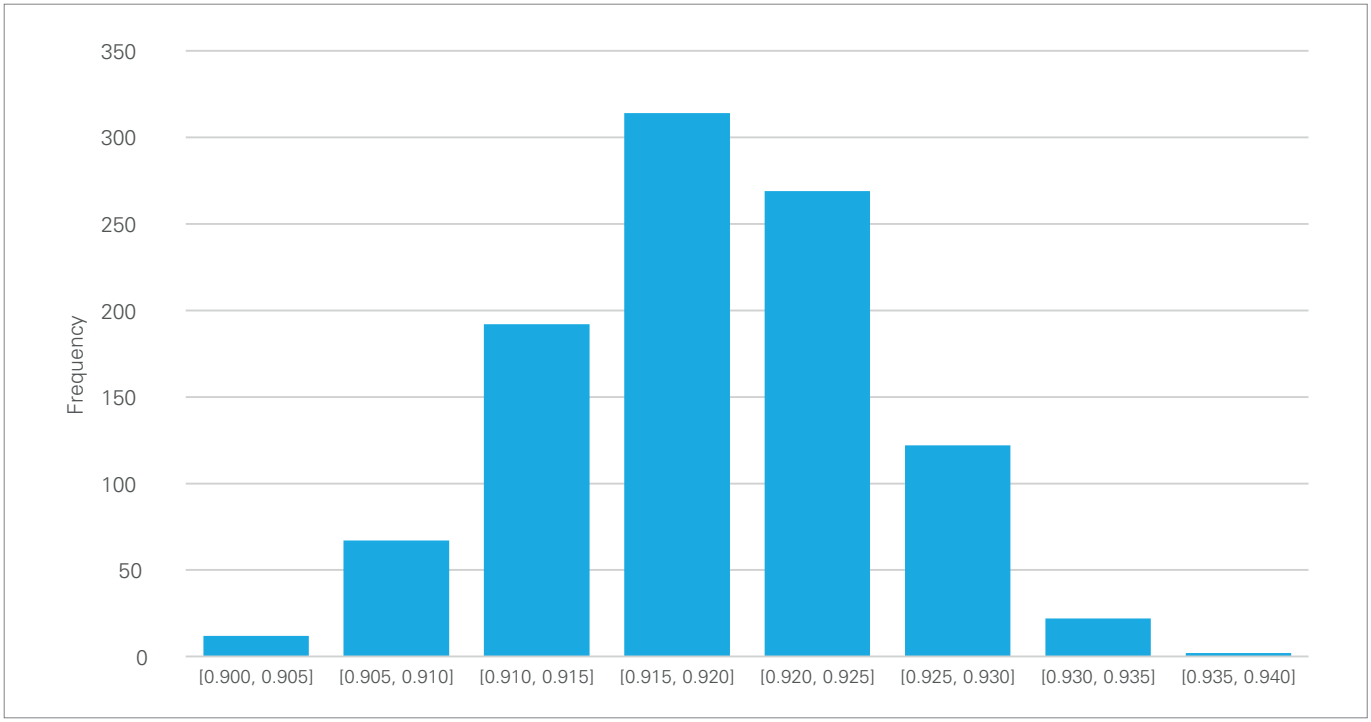


Figure 9: The base cost only UQ efficiency score in the PR14 Split is sensitive to measurement error

Source: Vivid Economics

The expected use of more disaggregated modelling makes the use of model diagnostics more important at PR19. As PR19 will involve subservice-level price controls in wastewater and water, it will be more important to understand fully the underlying causes of differences in model explanatory power, which may reflect inconsistent cost allocation or substitution of activities between services. The fact that bioresources price controls will not include enhancement spending means it is less likely that there will be significance outperformance in this area.

The pronounced effect of measurement error on efficiency scores underlines the importance of data diagnostics. As set out in Section 5, key drivers are measured with significant error, which has two implications for scores. First, it causes attenuation bias that can reduce the efficiency scores of companies with relatively high explanatory variable values. This will affect the magnitude of an efficiency challenge and will mean it applies to some companies more stringently than others. Second, it means that more unexplained variation can be attributed to factors other than efficiency. The Monte Carlo simulation described in Section 5.2 finds efficiency scores to be highly sensitive to measurement error in datashare variables. Figure 9 shows the distribution of the upper quartile efficiency score over 1,000 simulation runs on the PR14 split, finding that measurement error in a subset of drivers can account for more than a quarter of the overall challenge, equivalent to a range of £590m in totex in AMP7.

A more complete account of measurement error could inform the efficiency challenge at PR19 by improving the understanding of unexplained variation. This could include an assessment of measurement error in costs and non-datashare variables, which is not covered in the Monte-Carlo analysis.

6.3 DYNAMIC CHALLENGE

This section considers how modelling evidence on cost trends could be interpreted when setting a dynamic challenge. It considers how dynamic efficiency, reflecting technological improvement and learning, might be understood separately from omitted cost drivers which trend over time, potentially including quality drivers, and real price effects (RPEs). By better understanding the magnitude and direction of quality and RPE trends, Ofwat can set a more accurate dynamic challenge at PR19.

PR14 wastewater model coefficients imply that the level of efficient costs increased over time. The linear time trend included in all wastewater models *had positive* coefficients, implying efficient costs rise in real terms over time. This does not entail that companies grew less efficient over time in the historical data. Omitted or partially captured drivers such as quality, as well as RPEs are likely to have contributed to the positive trend observed at PR14.

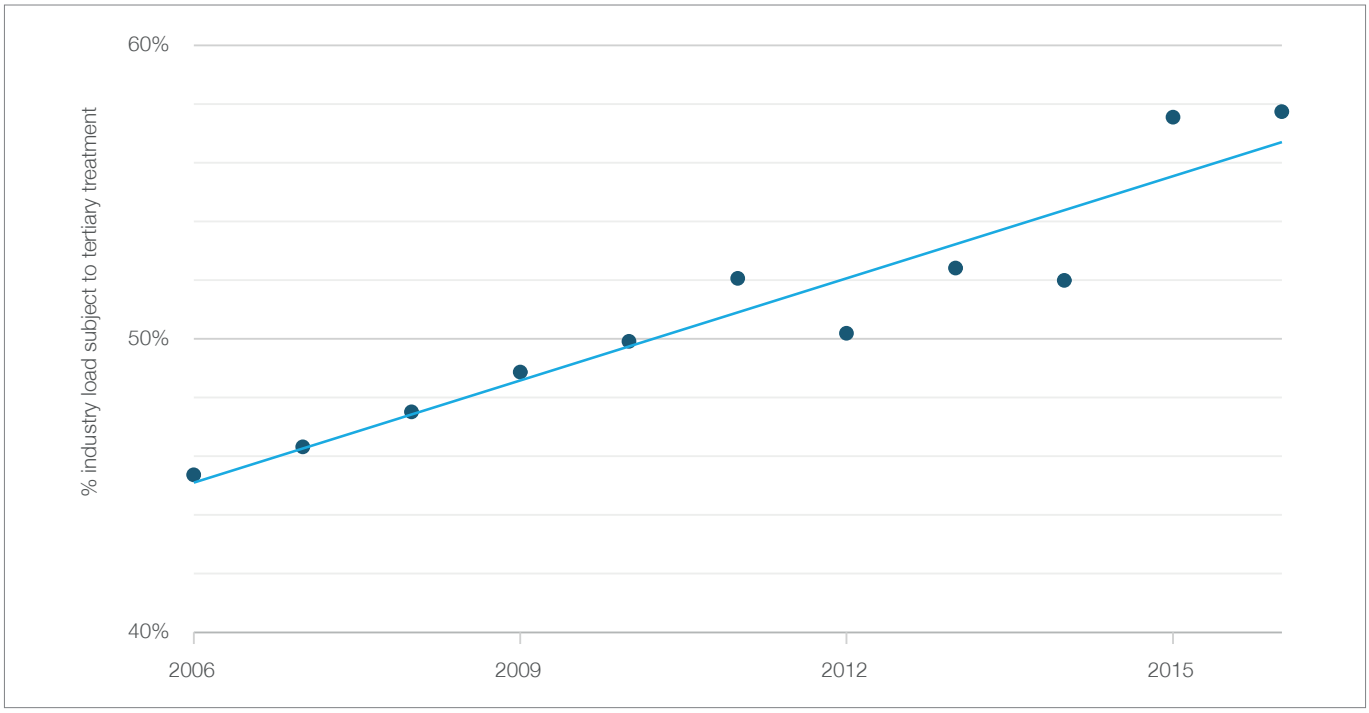


Figure 10: Tertiary treatment levels are rising over time across the industry

Source: Vivid Economics

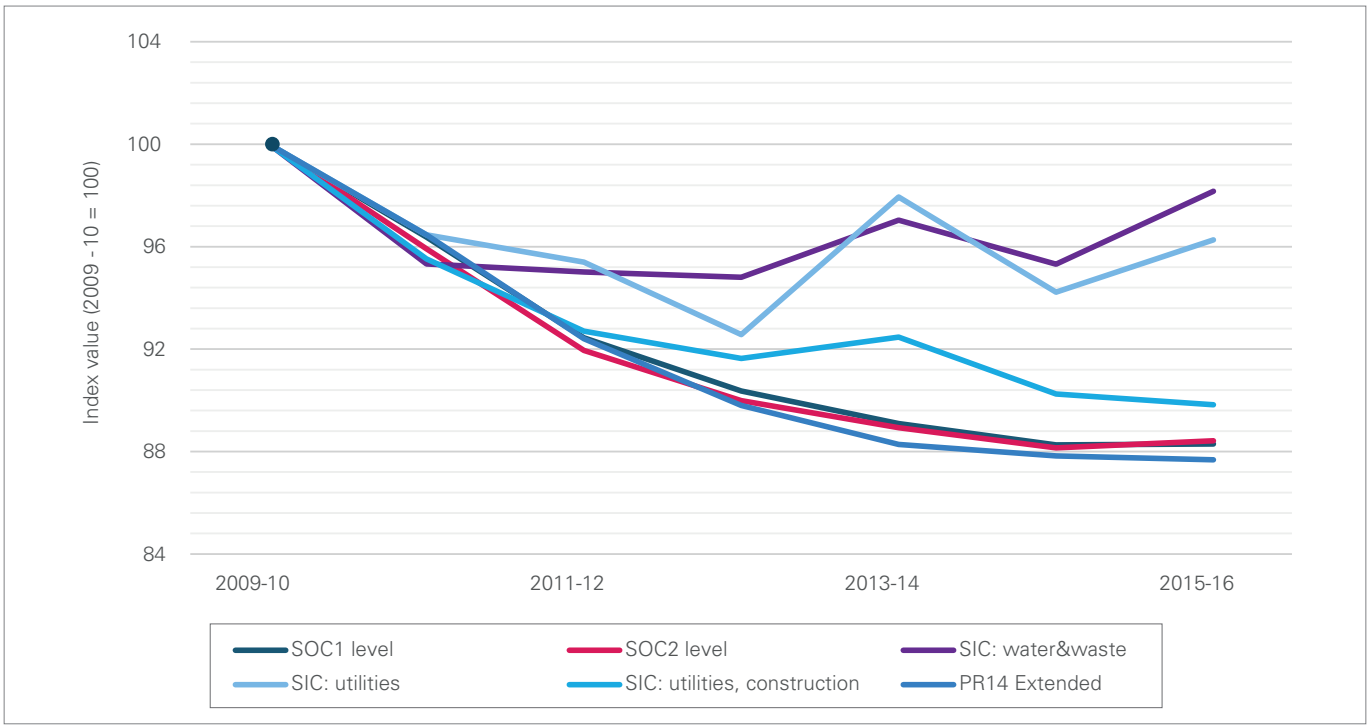


Figure 11: Occupational and industrial category based wage indices grew at very different rates in the historical data

Note: Real wage indices are based on data from ONS, defined fully in the June 2017 report. The PR14 index is the index used as an explanatory variable in the PR14 econometric models; SOC1 and 2 are wage indices based on representative Standard Occupation Codes from across the sector; SIC indices are wage indices from different Standard Industrial Code sectoral groupings.

Source: Vivid Economics

Service quality is a driver of costs that was omitted from PR14 econometric models and trends upwards. As shown in Figure 10, tertiary treatment levels have been rising in the industry across the last three AMPs. While tertiary treatment is an imperfect proxy for quality as discussed in Section 4.3, a stable trend across the entire industry suggests that treatment quality is probably rising over time. Tertiary treatment is associated with higher unit costs than other forms of treatment, as set out in the June 2017 Report. At PR14, the omission of treatment quality drivers is likely to have contributed to the positive time trend in econometric models. Similar trends are expected in network and bioresources activities, with companies meeting increasingly stringent environmental regulations and achieving reductions in flooding.

RPEs lead to time variation in real costs, but historical trends are not likely to be representative of conditions in AMP7. RPEs occur when company input prices grow at a different rate to the inflation index used to rebase their costs. If RPEs are left uncorrected, they can be conflated with changes in efficiency over time. This is particularly problematic if historical RPEs are a poor indicator of future RPEs so historical trends in costs are unrepresentative of future trends, as is likely to be the case for AMP7 where Brexit may affect input costs significantly. Ofgem's approach in the 2014 ED1 price control, where historical RPEs were projected forward on an input-by-input basis, was criticised for being backward looking (CEPA, 2014). By taking future RPEs into account when setting the dynamic challenge, Ofwat can ensure that the level of challenge is appropriate in light of real price changes driven by macroeconomic conditions.

Appropriate RPE adjustments use sector-specific price indices, but these can be sensitive to the choice of index. Figure 11 shows that regional wage indices based on industrial and occupational categories grow at very different rates in the historical data. Wages are a major component of total costs, so the choice of wage index will have a significant impact on the size of the overall RPE adjustment. Sensitivity testing of the RPE adjustment size to choices between input indices can help ensure a robust adjustment to time trends in modelled data.

Finally, understanding company asset and operating conditions can improve Ofwat's understanding of the future scope for efficiency gains. Where companies operate assets with headroom, productivity will grow as populations increase, but if new capacity is needed, this will tend to reduce productivity growth. Therefore, it is important to understand the extent to which historical growth has been accommodated through existing headroom or new capacity and whether this differs from expected future patterns.

6.4 RECOMMENDATIONS

Percentile challenges such as the upper quartile challenge are not robust to changes in models and data: such a challenge cannot be adopted as an efficient benchmark until model results have been finalised. Analysis shows that the magnitude of an upper quartile challenge varies between equally valid modelling suites, with the benchmark varying by 10 percentage points between base costs models despite no underlying difference in relative efficiency. The most appropriate level and form of challenge depends on the final suite of models adopted in PR19.

Diagnostic testing of models and data can inform the choice of efficiency challenge. Diagnostic tests reveal the explanatory power of the models and whether, as in the case of enhancement, the modelling approach offers a biased estimate of performance. Monte-Carlo simulation can highlight the degree to which efficiency scores and percentile challenges are affected by measurement errors and can indicate the extent of attenuation bias, which tends to exaggerate efficiency score ranges and penalise companies with 'large' driver values.

Modelling evidence can be grounded on an explicit view as to a reasonable level of efficiency variation. This view would not be expected to differ greatly between wholesale water and wastewater or between more granular subservices.

Evidence on companies' ability to meet the PR14 efficiency challenge should account for the way this were derived from models. Approximately half of the 10.5% efficiency challenge applied at PR14 reflected outperformance in enhancement, where 8 out of 10 companies outperformed their modelled allowances in aggregate. This arose because companies with larger programmes in any particular area had lower unit costs – so companies' costs tended to be concentrated in areas where models made them appear most efficient. Without this quirk in the modelling approach, the 10.5% challenge would have been substantially more difficult for companies meet.

It is sensible to set explicit dynamic efficiency challenge at PR19 rather than simply projecting past trends forwards. Better understanding the contribution of quality and Real Price Effect trends to time variation in costs will improve the chances of successfully implementing an appropriate and stretching challenge over the entire AMP.

Appendix

REGRESSION RESULTS

	NETWORK BASE	NETWORK AND ENHANCEMENT
Total length	0.38	0.40
Annual runoff	0.31	0.26
2011/12 dummy	0.00	0.00
2012/13 dummy	-0.10	-0.06
2013/14 dummy	-0.03	0.06
2014/15 dummy	-0.03	0.11
2015/16 dummy	-0.06	-0.05
2016/17 dummy	0.01	0.03
% urban population	0.74	1.20
Constant	-2.35	-2.36
N	60	60
R^2	0.86	0.88
VIF score	5.47	5.47
RESET test	Fail	Fail

Table 11: Network models regression results

Note: Ramsay RESET test outcome based on F test using 5% significance level

Source: Vivid Economics

Key:
- Significant at 1%
- Significant at 5%
- Significant at 10%

	TREATMENT AND ENHANCEMENT	TREATMENT AND SLUDGE BASE 1	TREATMENT AND SLUDGE BASE 2
Total load	0.98	0.94	0.95
% load bands 1 to 3	11.72	9.59	9.52
2006/07 dummy		0.00	0.00
2007/08 dummy		0.06	0.06
2008/09 dummy		0.11	0.12
2009/10 dummy		0.15	0.17
2010/11 dummy		0.12	0.13
2011/12 dummy	0.00	0.10	0.12
2012/13 dummy	0.09	0.11	0.12
2013/14 dummy	0.07	0.08	0.10
2014/15 dummy	0.07	0.10	0.11
2015/16 dummy	0.12	0.11	0.13
2016/17 dummy	0.15	0.14	0.16
% urban population	2.17	2.06	2.02
% sparse		0.54	0.67
% tertiary treatment	0.25	0.15	
% advanced tertiary treatment			0.21
Constant	-10.03	-9.17	-9.16
N	60	110	110
R^2	0.91	0.96	0.96
VIF score	2.59	2.42	2.39
RESET test	Fail	Fail	Fail

Table 12: Treatment model regression results

Note: Ramsay RESET test outcome based on F test using 5% significance level

Source: Vivid Economics

Key:
- Significant at 1%
- Significant at 5%
- Significant at 10%

	BIORESOURCES 1	BIORESOURCES 2
Sludge produced	1.01	1.03
% load bands 1-3	5.42	5.46
% sparse		0.14
Constant	-1.38	-1.47
N	60	60
R^2	0.80	0.80
VIF score	2.42	2.18
RESET test	Pass	Pass

Table 13: Bioresources model regression results

Note: Ramsay RESET test outcome based on F test using 5% significance level

Source: Vivid Economics

Key:
- Significant at 1%
- Significant at 5%
- Significant at 10%

	NETWORK PLUS	BOTEX AND ENHANCEMENT
Total load	0.86	0.87
% load bands 1-3	10.33	5.80
2011/12 dummy	0.00	0.00
2012/13 dummy	0.09	0.04
2013/14 dummy	0.08	0.05
2014/15 dummy	0.06	0.07
2015/16 dummy	0.10	0.01
2016/17 dummy	0.11	0.05
% urban population	2.34	1.01
% tertiary treatment	0.23	0.30
Constant	-8.04	-6.69
N	60	60
R^2	0.94	0.95
VIF score	2.79	2.79
RESET test	Fail	Fail

Table 14: Botex and Network+ model regression results

Note: Ramsay RESET test outcome based on F test using 5% significance level

Source: Vivid Economics

Key:
- Significant at 1%
- Significant at 5%
- Significant at 10%

LANDBANK DATA ANALYSIS

ADAS DATA PROCESSING

The core dataset for the analysis is the landbank available for spreading sludge. Publicly-available datasets do not allow a suitably-detailed and accurate analysis of net land bank availability. A proprietary dataset and tool, the Agricultural Land and Organic 'Waste': A National Capacity Estimator or ALOWANCE™, developed by RSK-ADAS Ltd was obtained. This is a GIS-based tool that estimates the available agricultural landbank in England and Wales for recycling organic materials, based on a number of physical and practical constraints (including topography, water courses, protected areas, soil conditions and livestock) and legislative restrictions on organic material recycling. The actual datasets used in the analysis are listed below:

ADAS available agricultural landbank data as a 10 km x 10 km map grid. Data was provided for all available years, namely 2000, 2004, 2010 and 2015.

ADAS breakdown of available agricultural landbank into arable land and grassland. Data was provided for 2000, 2004, 2010 and 2015.

Industry data on PE for individual site was converted to tonnes of dry solids, using a multiple of 60g/h/d. This enabled analysis of available landbank available within a selected radial distance of treatment works and sludge treatment centres (STCs), and development of an annual series. In order to process the data and improve the acceptable comparative appropriateness of the analysis the following additional key assumptions were made:

Available landbank was assumed to be usable agricultural farmland.

1-in-3 year rotation of sludge spreading per hectare (Ha) to account for nutrient management requirements. In practice this was applied by assuming the land available per year is a third of the total available landbank.

An average sludge application rate for arable land of 4.5 TDS/Ha/yr.

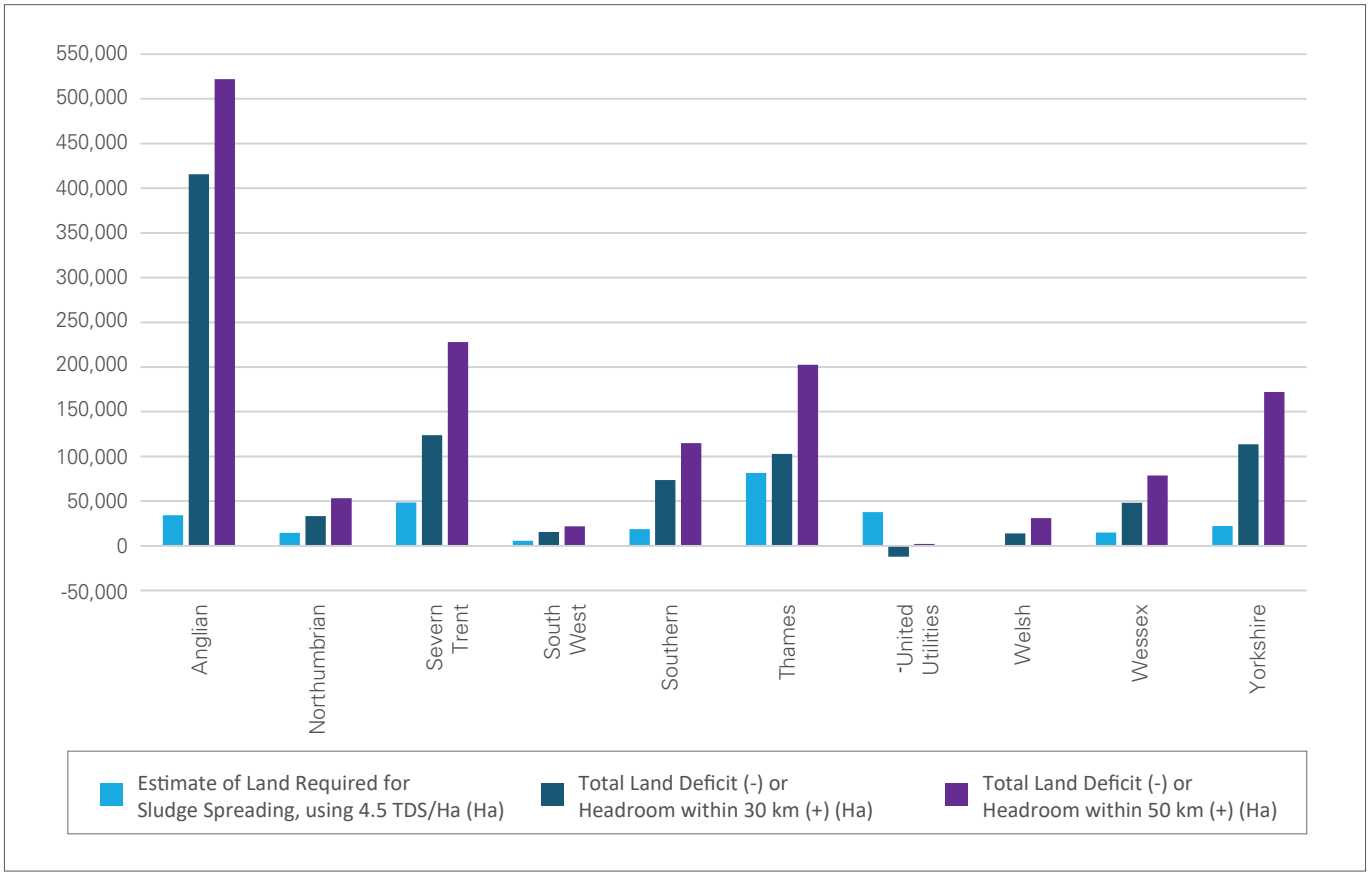
A conservative assumption that no spreading on of sludge was done on grassland. This results in conservative analysis that demonstrates the likely worst case. However, this assumption was applied equally across all WaSCs; the variation in grassland available from one company to the other (varying from a low of 15% to a high of 90%) is evident in the landbank deficits calculated.

The quality of sludge produced meets the minimum land disposal criteria i.e. conventional treatment, not advanced.

This analysis does not take account of temporary restrictions, such as the impact of rainfall or varying farmer uptake.

LANDBANK AVAILABILITY BY COMPANY

Sludge balances were estimated for radii of 30km and 50km, corresponding to distances cited by the Office of Fair Trading (OFT) as typical and maximum distances travelled by tanker for sludge disposal to land (OFT, 2011). Most companies have aggregate surpluses, but these vary significantly. United Utilities has an aggregate deficit within the 30 km threshold. For each company, sites with deficits were identified. Figure 12 below shows the deficits for each company plotted against TDS produced (Ha/TDS). Thames Water and United Utilities are most affected; Northumbrian Water and Severn Trent are the next most severely affected. The other companies have a modest deficit per TDS or none.



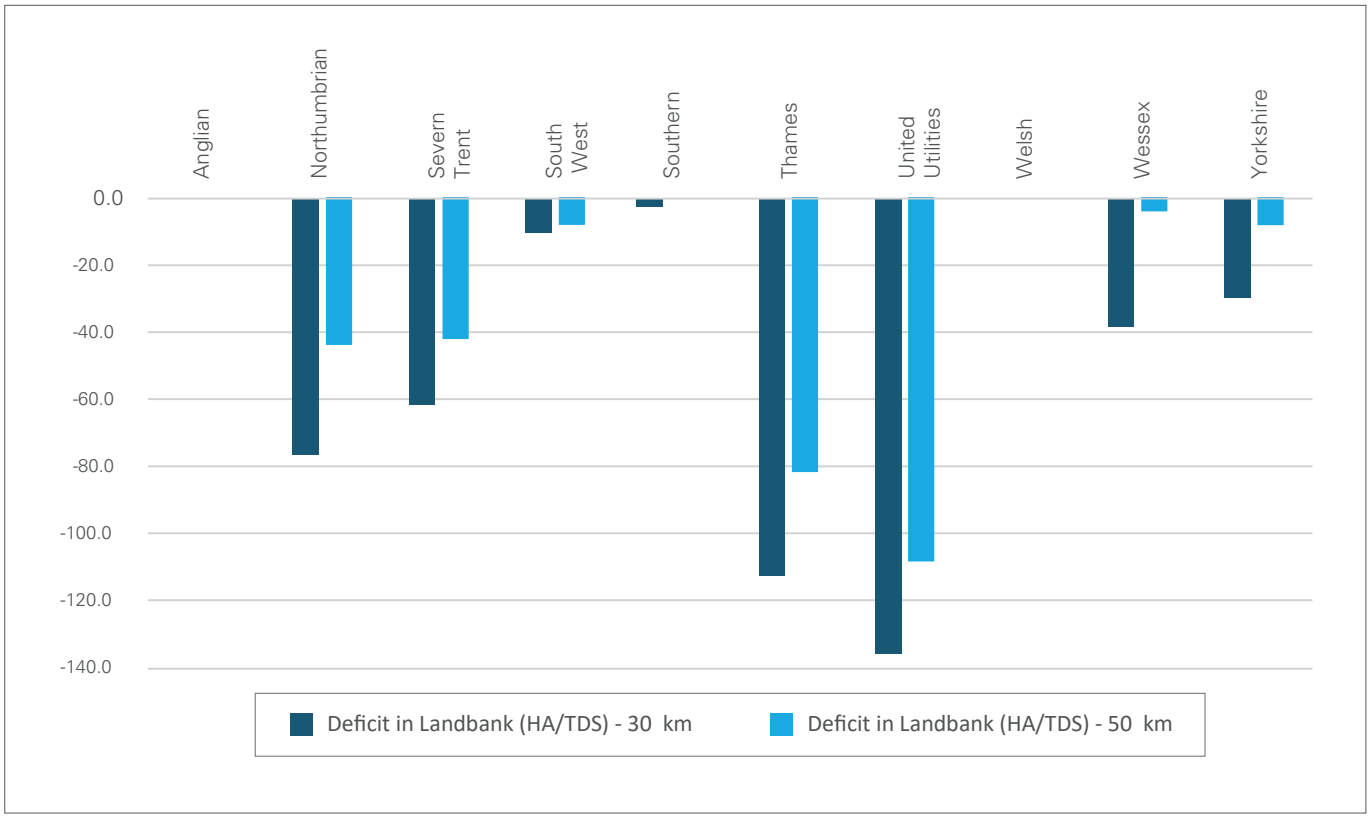


Figure 12: Land bank deficits within 30km or 50km radii
Source: Arup

OPTIMAL ALLOCATION OF SLUDGE TO LAND

The algorithm minimises the total sludge disposal distance, measured in tonne kilometres, associated with disposing sludge from Sludge Treatment Centres (STCs) or Wastewater Treatment Works (WwTWs) to available landbank. For each company, the following linear programming problem was set up: minimise sludge disposal distance subject to disposing of all sludge from each STC or WwTW, and total disposal to each cell of landbank satisfying the landbank's sludge receiving capacity. A computational solver, cbc, from the COIN-OR project was used to solve the resulting ten linear programming problems.

The key assumptions and features of the model were as follows:

Each Hectare of available landbank can receive a maximum of 1.5 tons of dry solids (TDS) per year (4.5 TDS / hectare divided by the crop rotation factor of 3).

Data on landbank availability was available for only 2010 and 2015, with a linear trend used to interpolate values for all other years.

The model's resolution was 10km x 10km: all STCs within each 10x10 cell were aggregated, and landbank availability was reflected as the percentage of each cell capable of receiving sludge.

The transport distance between any STC and landbank site was taken as the straight line distance between the two points.

Incineration or alternative disposal routes are not considered.

Although the algorithm can be run at higher resolution, for instance 5 x 5, the chosen level is considered to be sufficient granularity for modelling purposes. The use of straight line distances to reflect disposal distances for each route was a simplification. While road network layouts will result in 'true' optimal distances being greater than those identified here, the *relative differences between company optimal distances* are expected to be similar. This makes the current variable suitable acceptable for econometric modelling purposes. A more sophisticated model could be constructed to take road network layout into account when minimising sludge disposal distance.

Conflicts between companies over landbank availability were taken into account. The private optima for each of the ten companies can lead to multiple companies disposing to the same piece of landbank, and collectively violating the constraint of each landbank area. Figure 1 in Section 2.1 shows companies' first best allocations alongside each other, with conflicting demands for land highlighted in yellow cells.

To create company-level exogenous measures that are robust to conflicts where multiple companies seek to use the same areas of land, 'first best' and 'worst case' allocations were calculated for each company. In the first best case, the company in question could choose where to dispose of its sludge before any other company. In the worst case, all other companies chose where to dispose of their sludge before the company did. The mid-point between the two cases was then used as an exogenous measure. In practice, company choices of land in the 'first best' cases rarely came into conflict, meaning there was little difference in company-level transportation work between the first best and worst case scenarios. The mode implies that there is little competition between companies over landbank availability.

ENHANCEMENT MODELLING

Many of the enhancement models used at PR14 produce unstable estimates of unit costs that are dominated by outlying companies. For ten of the fourteen enhancement lines where unit cost models were used for cost assessment at PR14, it was possible to extend the sample of costs and drivers using more recent industry data. Table 15 presents a summary assessment of these extended models' performance. In most cases, updating the models led to implausibly large swings in estimated unit costs: for example, this was more than 20% for United Utilities in five of the seven cases where this could be calculated. For some lines, this stemmed from the single-driver models' inability to explain variation caused by a multiplicity of factors, reflected by R2 values substantially less than 0.8 or more seen in base cost models. For other lines, models produced high R2 values but only as a consequence of the models fitting the costs of a small number of firms that accounted for most spending in the area. In these cases, dominance of a few firms also means unit costs are unstable. Six of the ten models failed the Grubbs test, suggesting the presence of outlying companies with undue influence on model coefficients and forecast allowances.

COST LINE	EXPLANATORY VARIABLE	% CHANGE IN UNIT COST FROM PR14[1]	GRUBBS TEST	TOP COMPANY % SPEND	BOTTOM 5 COMPANIES % SPEND	LOG MODEL R^2	LINEAR MODEL R^2
First time sewerage	Properties	+38%	Fail	ANH – 46%	5%	0.45	0.86
Event duration monitoring (EDM)	Sites	N/A	Pass	UU – 20%	27%	0.28	0.42
UID storage	Storage volume (m^3)	>200 million %	Fail	UU – 86%	3%	0.57	1.00
Groundwater protection	Population served	N/A	Fail	SRN – 73%	2%	0.97	0.97
P removal	Population served	+6%[2]	Fail	SVT – 37%	13%	0.35	0.80
Sanitary determinands	Population served	-6%[2]	Pass	TMS – 31%	8%	0.68	0.72
UV disinfection	Population served	+74%[2]	Fail	UU – 51%	6%	0.98	0.59
Treatment growth	Population served	-55%	Pass	TMS – 30%	13%	0.26	0.76
Sewer flooding	Properties	-22%	Pass	TMS – 24%	23%	0.51	0.50
Odour	Complaints	N/A	Fail	TMS – 58%	12%	0.12	0.40

Table 15: Waste enhancement models based on the 16/17 datashare perform poorly on a variety of tests

Note: [1] Calculated for UU for 2016/17 datashare
[2] Denotes use of 2015/16 datashare

Source: Vivid Economics

Figure 13 sets out a candidate framework for assessing the appropriate modelling approach for individual enhancement lines. Inclusion in base cost models is tested first, for enhancement lines for which costs are allocated consistently across the industry. If activity is not substitutable with botex *and* expenditure not closely related to base cost drivers, area-specific models are considered. When appropriate drivers are not available, or model performance is affected by the presence of outliers, inclusion in botex models is again tested if lines are immaterial. For material lines which cannot be modelled directly or through the use of proxies, special factors or a dashboard assessment may be appropriate. Enhancement lines which cannot be modelled *and* are not material enough across the industry to warrant special factor claims are included under the unmodelled allowance.

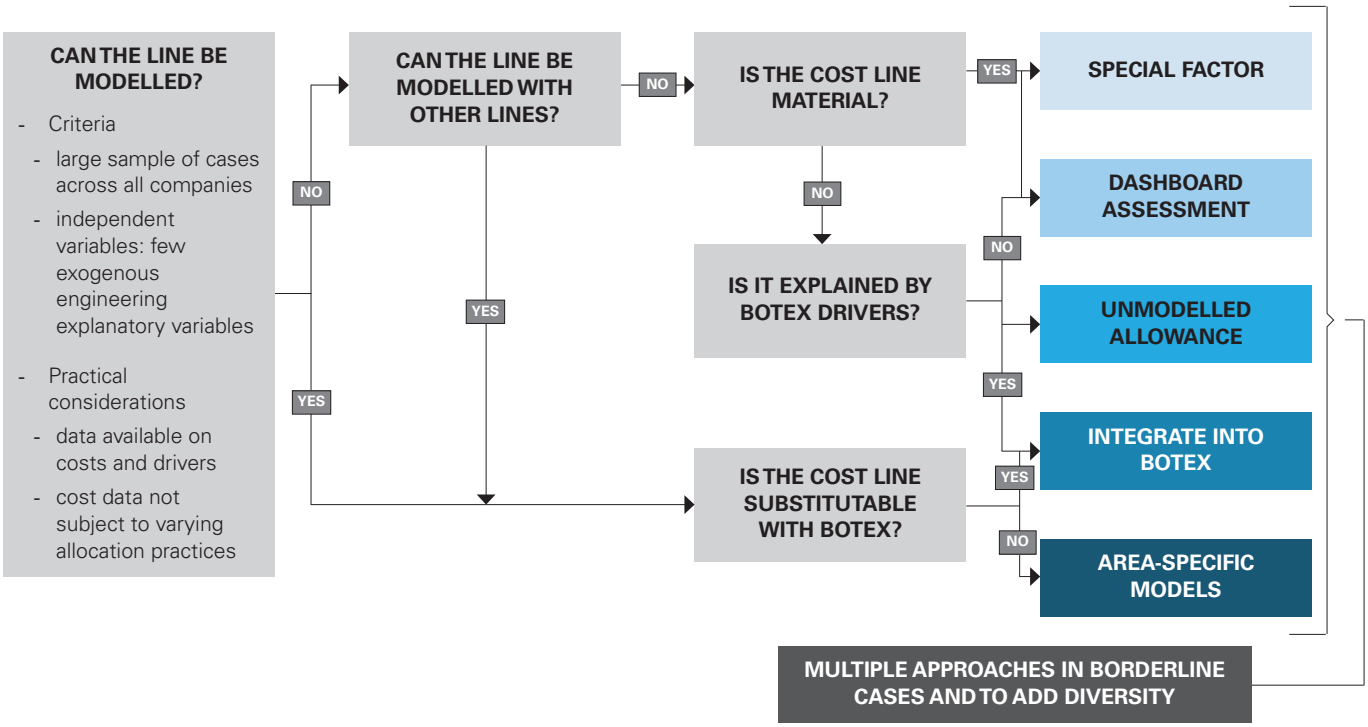


Figure 13: A formal enhancement modelling framework may improve modelling outcomes at PR19

Source: Vivid Economics

TRIANGULATION

The use of non-equal model triangulation weights can improve predictive power of suites of models. Models and totex split approaches which are equally well founded based on engineering and statistical criteria may make different contributions to overall diversity. For instance, an 'aggregated' totex split may contain more noise than a 'disaggregated' split, suggesting a lower weight for the former might be appropriate. An alternative to equal model weighting, mean square error optimal weights is set out below.

The mean square error (MSE) optimal weights approach minimises the variance of the combined forecast error term. High levels of noise in econometric models reflects uncertainty in forecast allowances and efficiency scores. An approach which minimises overall noise will therefore reduce price control risks for companies. Triangulation could be required at a number of levels at PR19, for instance at the subservice or service model level, and at the totex split level. The MSE approach could be applied at each level based on the combined forecast error term described below. The equations below set out the combined forecast error term, and the objective function for MSE optimal weights in the two model case:

$$e(\omega) = \omega\, e_1 + (1\text{-}\omega)\, e_2$$

$$\omega^* = \min_{\omega} \mathrm{Var}\,(e(\omega))$$

To avoid outlier companies having undue impacts on model weighting, residuals from log-models were used, rather than monetary values of unexplained cost. If model residuals were perfectly uncorrelated, model weights would reflect only the relative precision of each model.

Although MSE-optimal weighting has theoretical basis, it omits critical features of model diversity which make results difficult to interpret. The weighting methodology assumes that lower variance in the residuals reflect less modelling noise. Unexplained variation in econometric cost models reflects both modelling noise and efficiency differences. As MSE weighting cannot differentiate between these two factors, it is limited as an approach to model triangulation. The approach would overweight a mis-specified, overfitted model relative to a correctly specified model due to lower unexplained variation in the former.

The potential for negative weights is another drawback of the MSE optimal weights methodology. Negative weights arise when individual model residuals are strongly positively correlated. The combined prediction error is then minimised by offsetting individual model errors. Negative model weights potentially *increase* company risks, as triangulated allowances are not guaranteed to lie between individual model allowances. Similar residual values motivate the use of *equal* model weights, as model predictions are found to be mutually consistent.

The triangulation approach proposed by Anglian Water (Anglian Water, 2017) suffers from limitations similar to those identified above. The quality-adjusted triangulation approach is backward-looking, and will overweight overfitted models, as both R^2 and the Akaike Information Criterion reflect the degree of unexplained variation. The choice of criteria is also not justified through the underlying principles of diversity.

Equal model weights is a simple and transparent approach to triangulation which was used at PR14, and within other regulated industries in the past. It is a *neutral* choice, for companies and consumers as it does not over- or underweight models which individual companies perform particularly well or poorly in.

Triangulation based on equal weights is an acceptable approach given the limitations of formal alternatives such as MSE weights. Equal model weights is a simple, transparent and *neutral* approach which has regulatory precedent, and does not over- or underweight models which individual companies perform well or poorly in. As triangulation is only necessary when there is *more than one model which is considered best* based on available engineering and statistical evidence, equal weights are justifiable. More principled alternatives to model triangulation are difficult to apply and do not currently represent credible alternatives.

However, caution is required when interpreting results and setting out the implications of equal weights triangulation. While triangulation reduces the noise from any individual model, assuming a simple average between model estimates eliminates modelling noise is incorrect. Model noise can have unequal effects, with greater variability in allowances from alternative modelling approaches for some companies than others.

STOCHASTIC FRONTIER ANALYSIS

Panel data approaches, such as Stochastic Frontier Analysis (SFA), do not offer a credible alternative to established methods. Unlike Ordinary Least Squares approaches, SFA allows model residuals to be decomposed into company specific efficiency and white noise terms. In theory, this permits better estimation of relative efficiency differences, as the extent of modelling noise and biases is taken account of. However, the procedure is not robust to omitted variable biases or one-off forms of measurement error. Such persistent effects would be reflected in company efficiency terms, rather than the noise term. The strength of the evidence shown suggests that efficiency approaches which continue to conflate these effects do not represent an improvement over the UQ efficiency score approach. SFA is also based on strong yet arbitrary distributional assumptions, and like PR14 random effects panel models, is found to be unstable in small samples such as wastewater and water cost modelling exercises.

LIST OF ABBREVIATIONS

ABBREVIATION	MEANING
AMP	Asset Management Period
ASP	Activated Sludge Plant
BOD	Biochemical oxygen demand
CSO	Combined Sewer Overflow
EA	Environment Agency
GLS	Generalised Least Squares
Ha	Hectares
OLS	Ordinary Least Squares
ONS	Office of National Statistics
PE	Population Equivalents
PR14	The price review undertaken in 2014
PR14+	The models or supporting time series dataset used at PR14, extended by three years to 2015-16 using Ofwat 2016 datashare
Ramsey RESET	Ramsey regression equation specification error test
TDS	Tonnes of dry solids
UV	Ultraviolet treatment (wastewater)
VIF	Variance Inflation Factor
WaSC	Water and Sewerage Company
WoC	Water only Company

CREDITS

PROJECT LEADS

Arup Project Director: **Ian Gray**

Arup Project Manager: **Philip Songa**

Vivid Economics Project Director: **Robin Smale**

Vivid Economics Project Manager: **Oliver Walker**

INTERNAL PEER REVIEWER

Robin Smale
Director, Vivid Economics

EXTERNAL PEER REVIEWER

Dr. Ralf Martin
Assistant Professor in Economics, Imperial College Business School

ACKNOWLEDGEMENTS

The project team wishes to thank all of the staff at Arup and Vivid Economics who have contributed to the report, as well as the peer reviewers and other parties who have kindly shared knowledge, supporting information and insight, including the project sponsors, United Utilities.

DISCLAIMER

This report takes into account the particular instructions and requirements of our client. It is not intended for and should not be relied upon by any third party and no responsibility is undertaken to any third party. The information included in this work, while based on sources that the authors consider to be reliable, is not guaranteed as to accuracy and does not purport to be complete. The authors confirm that auditing of calculations and estimates was not part of the peer reviewers' scope.

: vivideconomics ARUP