CEPA

Opex model assessment criteria

January 2019

Northern Ireland Utility Regulator

FINAL

IMPORTANT NOTICE

This report was prepared by Cambridge Economic Policy Associates Limited (CEPA) for the exclusive use of the Northern Ireland Utility Regulator (UR).

Information furnished by others, upon which all or portions of this report are based, is believed to be reliable but has not been independently verified, unless expressly indicated. Public information, industry and statistical data are from sources we deem to be reliable; however, we make no representation as to the accuracy or completeness of such information, unless expressly indicated. The findings enclosed in this report may contain predictions based on current data and historical trends. Any such predictions are subject to inherent risks and uncertainties.

The opinions expressed in this report are valid only for the purpose stated herein and as of the date of this report. No obligation is assumed to revise this report to reflect changes, events or conditions, which occur subsequent to the date hereof.

CEPA does not accept or assume any responsibility in respect of the report to any readers of the report (third parties), other than the client(s). To the fullest extent permitted by law, CEPA will accept no liability in respect of the report to any third parties. Should any third parties choose to rely on the report, then they do so at their own risk.

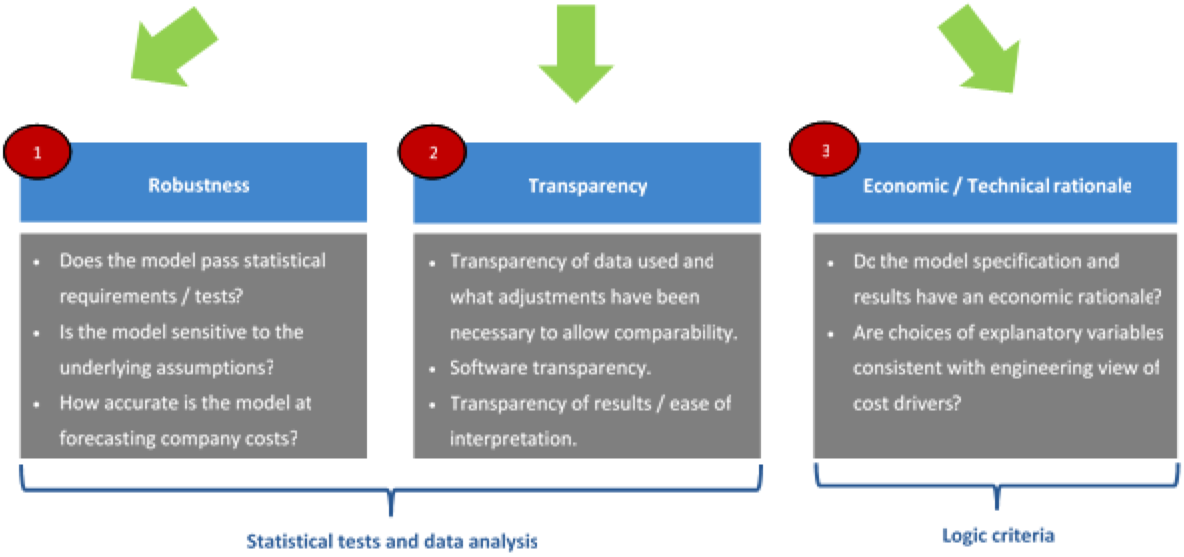
CONTENTS

1. INTRODUCTION

As part of model development, it will be important to have a clear process that allows us to evaluate the robustness of the models. This note setscriteria to guide this selection. Thes high-level criteria, summarised in the figure below, will form the basis of CEPA deciding whether the models are sufficiently robust to be used in UR's opex efficiency assessment for PC21. The proposed criteria are consistent with those used by Ofwat, UR and other regulators in developing cost assessment models. The proposed criteria apply specifically to the econometric models and should be considered within the overarching modelling strategy, set out in a separate paper.

When assessing model robustness, the criteria can be summarised into three high level areas that are presented in the figure below:

Figure 1.1: Model selection criteria



Source: CEPA

For each criterion, this note details:

- the specific tests CEPA will perform (and their interpretation); and
- how the tests will inform the model development process.

It is important to note that when several models are shown to have the right characteristics, UR may elect to use a reduced number of these models. When selecting UR would need to consider trade-offs between models e.g. one model may have a more theoretically correct cost function while another may be more parsimonious and have more intuitively appealing coefficients. Selection based on these trade-offs is not part of this paper but the results of the different tests described below will be an important input to the decision-making process.

2. MODEL ROBUSTNESS

Robustness of econometric benchmarking models can be interpreted in many ways and it can be assessed using different tests. We split these tests into three different categories:

- **Statistical robustness:** These tests focus on assessing whether these models have the right statistical characteristics.
- **Sensitivity of models:** These tests focus on evaluating whether the model results (e.g. coefficient estimates or efficiency scores / rankings) are robust to changes in the underlying assumptions and data.
- **Explanatory power of the model**: These tests focus on assessing the capacity of the models to explain the data.[1]

Each of these areas, and their associated tests, are discussed below.

2.1. STATISTICAL ROBUSTNESS

The tests that CEPA is proposing to introduce to evaluate whether a model has the expected statistical characteristics can be categorised into two broad areas, which are discussed below:

- Statistical significance of explanatory variables; and
- Testing the underlying assumptions of the models.

2.1.1. Statistical significance of explanatory variables

The statistical significance of a coefficient indicates the precision of estimates and whether we can confidently say the impact of an explanatory variable is not zero.[2] Ideally, we would not include insignificant variables but there may be trade-offs in terms of having a more holistic model that protects against omitted variables versus statistical significance (discussed below).

We can formally test the statistical significance of explanatory variables using three test statistics related to the statistical significance of individual variables, groups of variables and the correlation between variables. These are summarised in the table below.

Table 2.1: Statistical significance of variables

Focus of test	Test statistic	Description
Statistical significance of individual parameters	T-test	A coefficient is significant when it can be tested that it is different from zero with a certain probability.
Statistical significance of multiple parameters	F-test	Tests whether a group of coefficients are jointly different from zero with a certain probability.

[1] In general we would expect a model that fits the data well to also predict well. However, this is not necessarily the case if, for example, there have been changes in the way in which variables move together over the sample period and the forecast period.

[2] It is important to note that just because a coefficient is not statistically significant does not necessarily mean it is zero.

Focus of test	Test statistic	Description
Correlation between explanatory variables	Max or Min Variance Inflation Factor (VIF) >10	Models with a max and/or mean VIF above 10 are considered to have a relatively high risk of suffering from multicollinearity, i.e. some of the variables are correlates and providing similar information into the model.

Source: CEPA

When considering statistical significance, it is important to consider the potential distortion of analysis that could be generated by multicollinearity. Multicollinearity arises when two or more variables aim to explain the same information, e.g. km of mains and number of consumers are both likely to be higher for a large company, and larger companies have higher costs. Therefore, if variables with high multicollinearity are used in the model, the parameters will be picking up the same information. As a result, the effect will be divided between different cost drivers which could cause some variables to not be significant even when they reflect a relevant cost driver.

We propose to use a two-stage approach to minimise the risk of multicollinearity. First, we will not include any two variables in a model which are more than 90% correlated with each other. Secondly, we will also consider the VIF for each model. Even when a model cannot "fail" this test, it is standard practice to consider that models with VIF>10 (mean and max) could present multicollinearity issues.

2.1.2. Testing the underlying assumptions of the models

Econometric models rely on a number of underlying assumptions to produce consistent and unbiased results. These assumptions vary depending on the estimation method used. Furthermore, relevance to the robustness of the model for each of these assumptions varies significantly.

The table below sets out the tests we will use to determine if fundamental assumptions of the models have been broken and, for some assumptions, our approach to mitigating the impact of violated assumptions.

Table 2.2: Tests of underlying assumptions of models

Focus of test	Test statistic	Description
Linearity of the model	Ramsey's RESET test (F-test)	Tests whether there are any omitted non-linearities in the model. It can assist in choosing between Cobb-Douglas and other function forms.
Homoscedasticity of model residuals	White's test	Tests whether the error variance is constant across observations. To account for the fact that the variance between observations coming from one company and those coming from different companies could be different, we will use cluster robust standard errors.
Normality of model residuals	Joint skewness & kurtosis test (F-test)	The test used to evaluate statistical significance is based on the assumption that residuals follow a normal distribution. This is more important in small samples because we cannot use central limit theorem, which states that for large sample sizes the sampling distribution of the estimator converges to normality.
Pooling test	Chow test	For each model, we test whether the true coefficients of the pooled OLS model are significantly different from the true coefficients of the same model run on each individual cross-section of the data (i.e. year, 1, year 2,

Focus of test	Test statistic	Description
		etc.). If this test fails this provides evidence that panel data analysis may not be appropriate.
Tests of pooled OLS versus random effects	Breusch-Pagan LM test	Tests whether the variance of the individual fixed effect is equal to zero. If the test result implies that this is not the case (CEPA used a p-value of less than 0.01) this provides evidence that random effects or other panel estimation methods may be more appropriate.[3]
Test of fixed versus random effects	Hausman test	Tests whether the estimated coefficients between random effects and fixed effects are significantly different. If they are, this indicates that random effects estimation maybe biased and inconsistent. If the p-value of the test result is less than 0.01 CEPA considered that random effects estimation may not produce unbiased and consistent parameter estimates.

Source: CEPA

2.2. Sensitivity of results

Given the purpose of the selected models is to arrive at an efficiency challenge (if appropriate) that will be applied to NI Water's base opex, it is important that the models selected are not overly sensitive to changes in the underlying data or model specification, as this could indicate that the results are unstable, imprecise or influenced by certain types of data.

We will test the sensitivity of results by modifying:

a) the length of the panel (removing years);

b) the companies included in the sample (dropping companies); and

c) modifications of the assumptions used when developing models.

In addition to evaluating the stability of the coefficients, it is important to consider the stability of efficiency estimates across models. Therefore, we will also consider the dispersion of the level of efficiency/inefficiency obtained (i.e. the efficiency rankings and scores across models). Significant variations in the dispersion could provide an indication that either the model does not account for certain characteristics of the industry (omitted variables), or that one or more of the companies has characteristics that make it significantly different from all other companies and their inclusion will distort the final estimation.

The table below outlines all the robustness tests we have conducted to test the sensitivity of results.

Table 2.3: Testing the sensitivity of parameter estimates

Focus of test	Test statistic	Description
Introducing quadratic components when the RESET test fails.	Statistical significance of the new variable	CEPA will consider whether the new quadratic variable being included was consistently significant across models and had the expected sign (i.e. negative).

[3] If the unbiasedness and consistency assumptions of random effects hold then Pooled OLS will also produce unbiased and consistent parameter estimates. The only benefit of random effects is that it takes into account the structure of the data and is therefore a more efficient estimator than pooled OLS.

Focus of test	Test statistic	Description
Removing one year from the sample.	Seemingly Unrelated Regression Estimation (SURE) to test whether coefficients are equal between models	CEPA will evaluate whether two different sets of data (i.e. the one including/excluding the additional years/companies/variable) produce estimated coefficients that are significantly different from each other.
Removing the most or least efficient company in the sample for each model.		

Source: CEPA

Table 2.4: Testing the stability of relative efficiency across models

Focus of test	Description
Stability of inefficiency range	CEPA will conclude that a model fails to provide a consistent efficiency range if the efficiency score of the least efficient company is outside of a range of +/- 5 percentage around the average efficiency score of the least efficient company across all selected models.
Stability of efficiency rankings	CEPA will consider efficiency rankings to be unstable if the most/least efficiency company is not in the top 3 most/least efficient companies of the average efficiency rankings based on all selected models.

Source: CEPA

2.3. THE EXPLANATORY POWER OF THE MODELS

Even when the models will not be used to forecast the company's allowances, it is important that they appropriately and sufficiently capture the key cost drivers. For this reason, it is important to consider the capacity the model has to explain the data.

Table 2.5: Assessing the predictive power of the models

Focus of test	Test statistic	Description
Overall goodness-of-fit (OLS models)	Adjusted R^2	Measures the extent to which the explanatory variables explain the variation the dependent variable and across time while adjusting for the number of explanatory variables.

Source: CEPA

3. TRANSPARENCY

In addition to considering the robustness of models, it is important that the model selection process is transparent. The UR has highlighted that transparency is a priority for PC21.

We will attempt to ensure transparency in a number of ways:

- **Using a standard and readily available statistical package to conduct the analysis**. We will use Stata to develop models and will replicate the final models in Excels so that UR can easily re-run the analysis on an annual basis.
- **Providing a clear description and rationale of the data used**. We will provide details of data sources and any adjustments we have made in our final report.
- **Ease of interpreting model results**. We will ensure that the models can be easily interpreted and are not overly complex.
- **Data sharing.** We will share the final consolidated dataset with UR so it can be shared with NI Water and other relevant stakeholders who can replicate and challenge the proposed models.
- **Sharing the Stata do files with UR**, which could potentially also be circulated to NI Water if deemed appropriate by UR.

Most of the points above are process related, rather than specifically related to the models themselves, but complexity will be considered as part of the model selection process. However, while we will aim to develop models that are as transparent as possible, we need to be careful that not too much emphasis is placed on simplicity and remind ourselves of the fact that we are attempting to model a complex technology and want to obtain a good representation of the cost function and a reasonable estimate of inefficiency.

4. ECONOMIC AND TECHNICAL RATIONALE

One of the significant criteria which we will place a large weight on when developing the cost assessment models is economic and technical rationale. Economic and technical rationale can be broken down into three key questions that we will evaluate during the model development process:

Figure 4.1: Economic and technical rationale assessment criteria

1. Are the selected explanatory variables in line with our a priori expectations of what should be important explanatory variables?

- Explanatory variables should make sense from a technical engineering perspective as well as an economics perspective.
- As set out in our modelling strategy paper, cost drivers will be tested in order of importance: (1) scale, (2) density, (3) system characteristics, (4) level of activity and (5) quality.

2. Are the estimated model coefficients consistent with a priori expectations in terms of magnitude and sign?

- Our expectations of the sign and magnitude of coefficients on explanatory variables are set out in detail in our March 2018 cost assessment report for Ofwat. If any new variable is included into the analysis, the initial assumptions will be carefully explained in our report.

3. Are the selected models consistent with policy in other areas of the price control?

- An overarching requirement for all models will be consistency with the price control framework. For example, costs that are being assessed seperately should be excluded from the models.

Source: CEPA

5. USING TESTS TO INFORM MODEL DEVELOPMENT

Ideally the final models that are selected would pass all of the model tests discussed in this paper. However, setting such a high standard could make it very difficult to develop any models at all. While we set high standards, passing all tests is very challenging in applied work. As a result, as part of this work it will be important to understand what a model failing a test means for its potential use in PC21.

Before disregarding a model because it fails some of the tests, it will be important to consider the effect that failing that test has and the limitations it could impose on the use of the model. Trade-offs between test results are an inherent part of model development, meaning that a failure of one test will not necessarily result in the rejection of the model. But if we consider on balance that there are significant concerns which mean the model is not robust, we would go back through our iterative process and consider model alterations.

In order to illustrate how we consider the relative importance of each one of the tests discussed above, the table below presents the potential effect of a model failing specific tests (with further detail in Appendix A). In this table we have classified the tests depending on their importance into:

Table 5.1: PC21 Model Selection Criteria Summary

Level of importance	Definition
Very high	Tests that when failed, would disqualify the model automatically.
High	Tests that when failed would raise serious concerns about using the model in PC21.
Medium	Tests that when failed would raise some concerns about using the model in PC21 but the model could be used with caution if it passes other tests.
Low	Tests that when failed would raise very limited concerns about using the model in PC21. The model could be used if it offers the right incentives.

Source: CEPA

Table 5.2: PC21 Model Selection Criteria Summary

Level of importance	Tests
Very High	• Jointly statistically significant (F-test) • Overall goodness of fit • Consistency with policy in other parts of the price control
High	• Consistency with a priori expectations of magnitude and signs of estimated coefficients • Stability of efficiency rankings • Stability of inefficiency range • Transparency of results / ease of interpretation

Medium	• Sensitivity to:
	  ○ removal / addition of a year
	  ○ the removal of the most / less efficient company
	  ○ introduction of quadratic
	• Statistical significance of individual parameters (t-test)

Level of importance	Tests
	• Pooling test
Low	• Multicollinearity tests • Linearity • Homoscedasticity • Normality • Test of pooled OLS versus random effects (Breusch-Pagan LM test) • Hausman test for fixed effects • Comparison with PC15

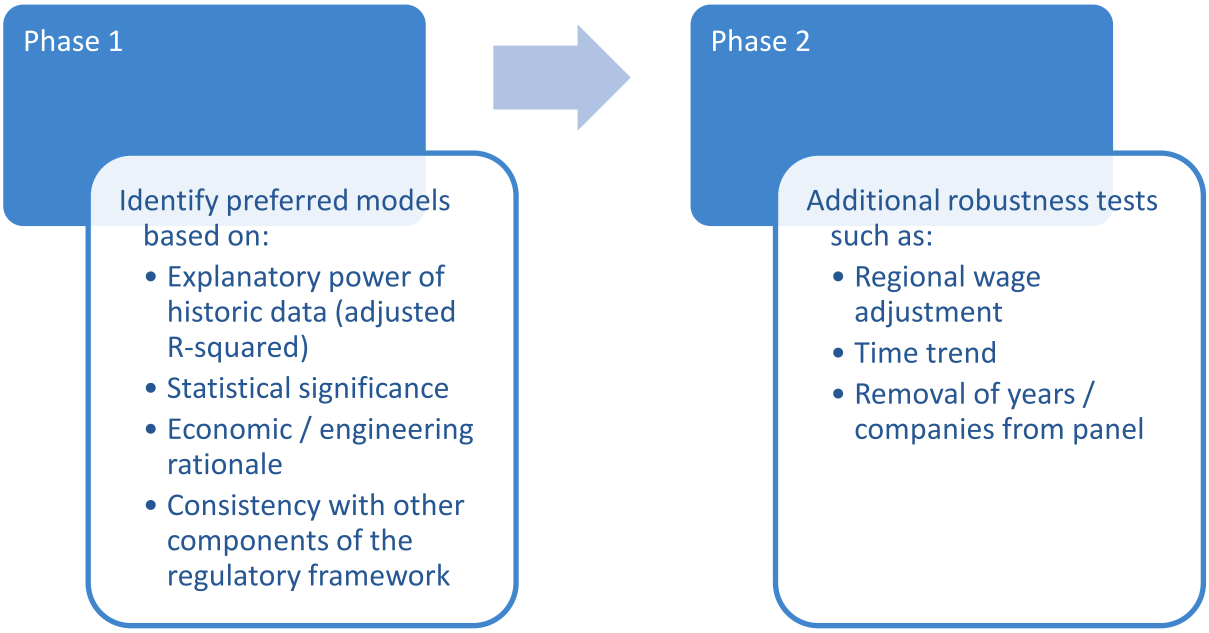
Source: CEPA

6. APPLICATION OF THE CRITERIA

To keep the modelling process manageable CEPA will carry out the analysis in two phases. In a first phase, CEPA will identify those models that meet the minimum characteristics required for a model to be considered further.

In a second phase, those models that are selected in Phase 1 will be evaluated further by running the remaining set of robustness tests discussed above.

Figure 6.1: Model development phases



Source: CEPA

APPENDIX A DETAILED PC21 MODEL SELECTION CRITERIA

Description	Level of importance	Comment
Robustness of models		
Statistical significance of individual parameters (t-test)	Medium	• If one or more of the coefficients in the model fails this test, we cannot rule out that the relationship being identified between the cost driver and costs under consideration is not spurious (i.e. the coefficient could be zero). • Parameters could fail this test because there is no relationship between the cost driver and the costs but also due to limitations in the data. The small size and poor quality of some of the components in the sample could make it difficult, if not impossible, to identify clearly the relationship between the variables and, therefore, we are unable to reject the null hypothesis that the coefficient is significantly different from zero. • While statistical significance of the estimated parameters is important, it is also important we can capture as many of the cost drivers as possible. This issue highlights the trade-off between parsimony and avoiding omitted variable bias, which is common in econometric modelling, but perhaps comes under greater scrutiny in the regulatory context. • As a result, it would be possible to include variables that are statistically insignificant if they reflect relationships that are well set in engineering and/or economic literature. In those cases we can be certain that the relationship exists even when there is not enough data or of enough quality to identify it robustly enough. • Furthermore, this would need to be compared with the F-test discussed below. Even when individual variables are insignificant, it is possible that they are jointly considering relevant effects. • One topic to be considered is whether this result is caused by the existence of multi-collinearity (i.e. high correlation between explanatory variables). If that is the case, one could decide to keep both variables but recognising that they are both measuring similar effects.
Jointly statistically significant (F-test)	Very high	• If the equation fails this test, it could suggest that the joint effect of all parameters is not statistically different from zero. • Therefore, if a model fails this test, it is not possible to determine whether there is an actual relationship between explanatory variables and the dependent variable. • There are different reasons that could justify this result (e.g. poor data quality or wrong specification of the model) but they all seem to indicate that there is a lack of statistical robustness that will make the result easy to challenge.

Description	Level of importance	Comment
Underlining assumptions tests		
Linearity	Low	• This test aims to determine whether one could expect a linear relationship between the cost driver and the costs under consideration. The linear assumption might be a reasonable assumption in some cases whereas in others it may not. • Failing this test seems to indicate that the data could be better fitted using a different functional form (e.g. quadratic). However, this is not to say that a linear assumption is automatically wrong but that other options could be better. The introduction of alternative functional forms, however, could increase the complexity of the models which would be linked to additional data requirements. • Given the need to develop transparent models and the limitations in the data available, the UR could still use models that fail this test. However, it will be important to consider whether additional adjustments need to be introduced in the results to account for the lack of linearity (e.g. introduction of quadratic terms or other explanatory variables).
Homoscedasticity	Low	• Ensuring that OLS is BLUE (Best Linear Unbiased Estimator) requires that the residuals of the equation are normally distributed with an average of zero and a variance equal for all of them. If this assumption is violated, the results are still unbiased although they could lose some other properties. Heteroscedasticity can be detected by inspecting the residuals in addition to formal testing procedures. • If a model fails the homoscedasticity test, it means that the variance of the errors is not equal for all observations. Different measures can be introduced to address this issue (e.g. use cluster robust standard errors). However, if the effect persists, the model could still be used as the results are robust.
Normality	Low	• The impact of non-normality only has implications in small samples. As the sample size increases, the sampling distributions are approximately normally distributed. This means we can apply standard inference based on asymptotic approximations, and as a result normality is not a great concern.[4]

[4] Even in small samples, the lack of normality only has implications for the inference of t- and F-test statistics and not the unbiasedness and consistency of parameter estimates.

Description	Level of importance	Comment
Tests of pooled OLS versus random effects models - Breusch-Pagan LM test for random effects	Medium	• Both OLS and Random Effects assume that the individual firm effect is uncorrelated with the regressors. Thus, the main difference between OLS and a Random Effect estimation is the assumptions that are made about the structure of the error term. • If the model fail this test, then OLS is unbiased but not efficient (assuming the aforementioned assumption holds). Therefore, we can still use OLS to produce unbiased parameter estimates.[5]
Hausman test for fixed effects	Medium	• If the unobserved fixed effects are uncorrelated with the regressors then both OLS and Random Effects estimation produce unbiased results. However, if the unobserved fixed effects are correlated with the regressors only Fixed Effects estimation produce unbiased results. • The Hausman test can be used to test whether the unobserved fixed effects are correlated with the regressors. If the difference in the estimated coefficients between Fixed and Random Effects estimation is statistically significant, this is evidence that the regressors are correlated with the unobserved fixed effects. In this case we will need to consider whether fixed effects estimation is more appropriate and/or whether there are any omitted but available time invariant explanatory variables we could test in the random effects model. • Nevertheless, while Fixed Effects estimation has useful statistical properties it is rarely used in efficiency analysis because of two reasons. Firstly, it is difficult to distinguish between inefficiency and company heterogeneity. Secondly, due to the relatively small datasets, fixed effects estimation tends to produce very wide standard errors. As a result, OLS or random effects estimation, while biased, is often preferred to fixed effects estimation within an efficiency analysis exercise.
Sensitivity of results		
Chow test - Sensitivity to removal / addition of a year / company	Medium	• This test would consider whether there is any data that does not fit with the rest of the data set (i.e. a company or a year presenting different characteristics than the rest of the data set). There are several reasons that could justify this distinction such as structural break in the data (different across years) or the presence of an outlier in the data. • Therefore, before taking a specific decision it will be important to evaluate the rationale that could justify these differences. For example, if a company has a very different cost

[5] Assuming the individual fixed effects are not correlated with the regressors.

Description	Level of importance	Comment
		structure than the rivals for, for example, historic reasons outside of the control of the company, it could require that that company is excluded from the analysis.
Sensitivity to inclusion / exclusion of quadratic terms	Medium	• This test considers whether a quadratic function form fits the data better than a linear functional form. Statistical significance and consistency across model specifications (i.e. sign and magnitude of estimated coefficient) will be considered when deciding if a quadratic term is included in the model.
Sensitivity to inclusion / exclusion of explanatory variables	High	• This test considers the potential effect on the efficiency rankings for a company or group of companies of including/excluding an explanatory variable. This would allow to identify whether the model produces consistent efficiency rankings/scores. • There are reasons that could justify these changes in efficiency rankings/scores. Therefore, the results of this test will need to be carefully evaluated.
Comparison with PC15	Low	• Where possible, we will compare efficiency results with PC15 efficiency results for NI Water. • However, it is important to note that companies could have changed their efficiency (become more or less efficient over time) which will be reflected in changes in efficiency scores/rankings. Therefore, even if NI Water have reduced costs over the course of PC15 this does not necessarily mean their relative efficiency gap with England and Wales companies has also improved if the latter have become more efficient over time.
Predictive Power		
Adjusted R-squared	Very high	• If a model fails to explain a significant variation in the costs of the industry, it would be inappropriate to use it for the estimation of the costs going forward (for models in log-terms the R-squared relates to the log of costs). Therefore, we would expect that only models with a high explanatory power should be used as the base of the cost assessment methodology (e.g. above 80%).
Transparency		
Transparency of results / ease of interpretation	High	• To facilitate their use during PC21, the models should be understandable and intuitive. However, there would need to be a balance between simplicity and complexity if the latter brings a significant improvement in the performance of the model.
Data availability	High	• To ensure that NI Water can challenge the models, it is important that they have access to the data used in developing the models. • To reduce this risk CEPA propose that the final consolidated data set could be shared with NI Water.

Description	Level of importance	Comment
Software transparency	Very High	• We will use Stata to conduct our econometric analysis, which is an internationally recognised standard software for applied econometrics analysis. • In addition, our final model selection will be replicated in Excel to ensure that the UR can re-run the analysis on an annual basis.
Economic and technical rationale		
Are any important explanatory variables omitted?	Medium	• By omitting important explanatory variables the model would fail to incorporate some of the drivers into the analysis. In some cases it will not be possible to incorporate these variables as the model already includes a significant number of cost drivers given the data available, or no robust variable has been found to cover this specific cost driver. • Engineering and econometric experts will be used to minimise this risk. However, if it were to arise this would be flagged and potential off-model adjustments would need to be incorporated into the results to account for these effects.
Consistency with a priori expectations of magnitude and signs of estimated coefficients	High	• Ahead of running a regression CEPA will have an expected sign for the coefficients. In some cases, the economic and technical literature will also be able to offer an expected size for the parameter. • Estimated coefficients that significantly differ from our *a priori* expectations of magnitude and signs could be a cause for concern. However, there are good reasons that could justify this effect. For example, the variable could be picking up some additional effect for which the explanatory variable is only an imperfect proxy. if any variable would fail this test, it would need to be considered carefully and a good explanation developed before putting forward the model.
Consistency with policy in other parts of the price control	Very high	• Models that produce coefficients that are inconsistent with policy in other parts of the price control would be automatically rejected, e.g. inclusion of costs that are dealt with in other parts of the price control.

Source: CEPA

CEPA

Queens House
55-56 Lincoln's Inn Fields
London WC2A 3LJ
United Kingdom

Level 20, Tower 2
Darling Park, 201 Sussex St
Sydney NSW 2000
Australia

CEPA Ltd
@CepaLtd