Carlos Pineda
carlos.pinedabermudez@thameswater.co.uk

12 January 2023

Daniel Mitchell
Principal Economist
Ofwat
Centre City Tower,
7 Hill Street, Birmingham.
B5 4UA

Assessing Base Cost at PR24: Econometric Models Submission 2023

Dear Daniel,

We welcome Ofwat's initiative for companies to submit proposals for base econometric models as part of PR24. This provides the opportunity to improve the current PR19 econometric models with the inclusion of alternative cost drivers that could improve on the explanatory power and performance of the current suite of models within Ofwat's modelling guidance.

We want to highlight in this letter our main findings across the different areas explored during the modelling process in water, wastewater, and retail.

In Water, our main focus is on:

- The use of average pumping head or capacity of pumping stations as better proxies for energy cost drivers than the number of booster pumping stations.
- Drivers of capital maintenance i.e. age of network, replaced or relined mains and leakage
- The capacity of reservoirs for use in the Water resources plus models.

For wastewater our focus is on:

- The impact of rainfall on models and the use of Population Density versions LAD and MSOA instead of Property Density.

- The development of wastewater network plus models.

For retail our focus is on:

- Addressing the impact of the structural break likely caused by the covid event on the transience driver; and
- Improving the deprivation drivers.

All these areas improve on the existing PR19 models.

Before introducing the insights learnt from the models proposed in more detail, we would like to mention that all the analysis, in water, wastewater and retail models is based on Random Effects Models (RE) estimated by GLS using cluster robust standard errors. We did not find any empirical evidence that suggest the use of Pooled OLS models, which shows the significant presence of an unobserved time-invariant heterogeneity element across the botex models for all companies in the panel dataset; therefore, we rule out the use of OLS and use RE econometric models[1]. All our wholesale models are run using the period 2011-12 to 2021-22 and use the new PR24 Botex Plus definition proposed in the Stata Code published by Ofwat.

We provide our insights in the following paragraphs starting with Wholesale Water, followed by Wholesale Wastewater and finally Retail.

Wholesale Water

In our water models, we are proposing the use of *Average Pumping Head* (APH) or the *Capacity* of Pumping Stations (per property or per main) as potential new cost drivers in Treated Water Distribution (TWD) and Wholesale Water (WW) aggregated models, instead of the Number of Booster Pumping Stations per main (NBS). We have found that the use of these new drivers improves the performance of the PR19 models. For example, in the TWD proposed models TMSTWD1-16 the R^2 ranges between [0.959 – 0.973] versus the current PR19 version R^2 of 0.957. We consider that the current driver, NBS, does not reflect accurately the engineering and operational link with botex costs. As mentioned in a CAWG[2], there are concerns across the industry regarding the confidence of the NBS driver and what it is actually representing. The current approach of using NBS in TWD (instead of *APH* or *Capacity*) can be misleading as a proxy to explain power costs, as this driver is not describing the characteristics of Pumping Stations and how much, on average, they can reach/cover according to the area where they operate. Moreover, the correlation between Density and NBS could be generating multicollinearity issues (e.g., swap in the sign on parameters; stability of parameters; more can be found on the TWD template).

In cases where *APH* cannot be used in the models, we believe that *Capacity* might better reflect the topography and consequently the conditions and needs that each Pumping Station faces instead of NBS. Furthermore, the standard errors of the cost drivers and in particular the ones

[1] Although, our templates show in the robustness check section some tests that only run with Pooled OLS models.
[2] Cost Assessment Working Group (CAWG) session 4th on the 7th of September 2021.

related to *APH* and *Capacity*, improve significantly relative to those in the NBS-based models, providing more confidence in the estimated parameters. For these reasons, we have used either *APH* or *Capacity* in place of NBS in each of our proposed models. More details, empirical evidence and explanation are provided in the TWD template. Similarly, our proposed models in WW reflect a statistically significant effect of *APH* or *Capacity*, particularly when Capacity per property is used, showing the robustness of the driver across different levels of cost aggregation. Among the models proposed in WW, in particular models TMSWW13-27, we notice that when a Composite Scale Variable (CSV) is used alongside with *APH/Capacity* and *age of the network*, the R^2 improves versus the current PR19 models.

In addition to the *APH* or *Capacity* drivers, we explore drivers related to capital maintenance (1: age; 2: length of mains relined & renewed) and output-service-cost link (3: leakage). Although it could be argued that these set of drivers are under management control, they could also be considered as "*drivers that are only endogenous in the long term as the risk of perverse incentives is lower*."[3] We explore these drivers mainly in TWD and some (e.g., age of network) in WW[4]. The *age of the network* is a relevant cost driver that links to capital maintenance costs. The driver we propose appears to add a significant impact on base TWD costs (see in TWD models TMSTWD1-2 versus models TMSTWD3-4, 11-12). In the WW aggregate models, it shows a strong statistically significant effect, improving the R^2 of these models.

Regarding the other two proposed drivers, *Mains Relined & Renewed* (R&R) and *Leakage* explored in TWD models, we consider that R&R can be considered as a long-term commitment in the industry considering the rate of replacement likely to be faced by the industry in the next decades. This is an important driver for outcomes such as leakage or supply interruptions. R&R is statistically significant across all the specifications explored and under different definitions of the costs (e.g., Botex Plus as in PR19 or Botex or the Botex PR24 proposed definition). For instance, models TMSTWD1-2 versus models TMSTWD5-6. With respect to *Leakage* this driver links to the quality of the service provided by water companies. *Leakage* could be considered as a long-term commitment driver as the industry moves forward to a reduction of 50% of leakage levels by 2050. *Leakage* has a clear impact on customer preference and its impact on base costs can be seen in the models proposed TMSTWD7-10.

It is very important to re-consider the use of these drivers in the base models. The models that include *Leakage* estimate a negative impact of the leakage coefficient on botex cost reflecting the appropriate regulatory incentive to reduce leakage. All the estimated effects of the driver yield this negative sign effect and improve the R^2 when compared to TMSTWD1-2 and the PR19 model. These results go in line with the proposed models we provide in the "*Assessing Base Cost at PR24*" consultation.

Models TMSTWD9-12 provide different extensions using a combination of *age of the network*, R&R or leakage alongside with *APH* or *Capacity,* providing more evidence on how the models

[3] Ofwat's Draft and Final Methodology, Appendix 9, p.12.

[4] More details are provided in the templates for WW and TWD.

could be improved through a higher R^2 (e.g., 0.973 model TMSTWD12). Lastly, models TMSTWD12-16 instead of using LAD density use MSOA density and show the robustness of drivers such as *APH*, *Capacity* or R&R as well as the robustness effect of density when measured in different ways. Using MSOA density can also improve the fit of the TWD models (see for example the R^2 of model TMSTWD1 versus model TMSTWD13) and overall performance of the wholesale water models. We believe that this is because the more disaggregated targeted measure of density through the Middle Layer Super Output Area better represents the effect of very dense areas than the Local Authority District (LAD) level.

We also propose a few models in Water Resources Plus (WRP). The proposed models TMSWRP1-2 include the Capacity of Reservoir (Ml) per Property. This driver captures capital maintenance and operating costs as well as management of reservoirs across water resources. The driver shows a positive effect, and it is statistically significant in these two versions of the models as well as in models TMSWRP3-4. We believe that the *Capacity of Reservoir per Property* is a significant driver that adds information in explaining base costs, as reservoirs are sensitive to maintain and operate for security reasons (e.g., regular inspection walks, and maintenance is required as stated in the Reservoir Act 1975). Lastly, we note that the Ln(wac) driver is no longer statistically significant in the PR19 model for WRP. To address this, we considered an alternative driver for the effects of water treatment complexity, based on a re-calculation of the weights that are assigned to the complexity bands by grouping the lower and upper levels of complexity with a simple average weight. With this adjustment the wac driver shows a statistically significant effect across the specifications in models TMSWRP3-4 alongside the *Capacity of Reservoir (Ml) per Property* yielding an increase in the overall performance of the models.

Wholesale Wastewater

We present ten models for Sewage Collection (SWC) to provide potential evidence on how the current models in PR19 could be improved. We think that the use of average property density as a cost driver (Properties/Mains) does not reflect the different levels of density faced by each company across the industry. This seems to be inconsistent with the current PR19 model SWC2 that uses LAD density and its square term, that suggest that the different levels of densities faced by companies are relevant to understand the variation of botex cost across the industry. We believe that Weighted Average Population Density LAD and MSOA are good complements to show the effect of density in the industry as its square terms become significant in the models differentiating in this way the different levels of density faced by companies (see for instance models TMSSWC 2,5,8 and 10 for the use of MSOA density, its robustness and performance). We believe that an initial set of complementary models for SWC are models TMSSWC1-2 that use a LAD and MSOA densities. We consider that population density is more beneficial to understanding operating costs than property density, for the simple reason that it is the wastewater produced (load) by people that generates the cost to operate. Variances in population density versus property density per sq km take into account not only the property type i.e., 1 bedroom starter homes versus maisonettes, flats or large multi bedroom houses where the occupancy number will be greater but also a bit more about the demographics i.e., areas may vary in terms of occupancy based on possibly house or location value where it is more common for single occupancy of homes versus multiple occupancy due to individuals personal circumstances. For example, in wastewater hydraulic modelling the approach is always based on occupancy (population density) as it is not possible to derive wastewater usage profiles from just a house (property) count.

Alongside these density variables, we have found a significant effect of the rainfall variables proposed in the dataset. This follows our suggestion as a potential new cost driver in our December 2021 "*Assessing Base Cost at PR24*" consultation response. In that response, we provided some suggestions on how this rainfall driver could improve the fit of the sewage collection models. When rainfall is included, the performance of the R^2 is improved. Moreover, models that include either LAD or MSOA densities with its square terms and rainfall are more robust than a model that uses average property density[5]. Regarding the different measures of rainfall, we consider that the effect of *urban rainfall* LAD seems to provide more compelling results when compared to the other two alternatives of measuring rainfall. For example, the standard errors are lower and the R^2 higher when using *urban rainfall* LAD as a cost driver than when using annual rainfall. Lastly, we introduce the potential use of a Composite Scale Variable (CSV) for SWC models. The results indicate promising results and improvements in the models R^2, but more discussion on the weights of the drivers used in the CSV calculation is needed. We provide more discussion in the SWC template.

Regarding the Sewage Treatment (SWT) models proposed, we provide potential alternatives to the current PR19 SWT models. Our models TMSSWT1-2 , provide an improvement on the R^2 when compared to the current PR19 models SWT1-2. This is because of the inclusion of Pumping Capacity per Main. Sewage Pumping stations capacity can provide a good level of insight into the operation of a Sewage Treatment Works (STWs). In general, STWs will either receive incoming flow via gravity, pumped flows, or a combination of the two. For those sites where the dominant flow is pumped, correlation between the operation of the pumping station and the STWs can be hugely insightful, typically this will involve looking at the terminal pumping stations only i.e., those pumping stations that outfall directly to the STWs. The pumping station capacity can be helpful because it provides insight on the total flow passed to the STWs and not just the treated flow. Total flow will pass through screens etc. capturing costs that might not otherwise be allowed for. Model TMSSWT1 improves the fitness of the model but struggles with the RESET-test. These proposed models' R^2 are higher than the current PR19 models, with an R^2 of 0.89 versus 0.85.

All the models proposed for Wholesale Wastewater Network Plus (WWWNP) follow our response to question 8 of the "*Assessing Base Cost at PR24*" consultation. The models presented in the WWWNP template are a continuation of the insights proposed in the consultation response, in particular with the use of *Total Load* as the main scale driver. Furthermore, we explore a CSV variable, but *Load* still provides a better performance for the models overall. Our proposed models provide an R^2 that ranges between [0.907 – 0.963] depending on the model specification. Our

[5] See for instance RESET-test.

first two models TMSWWWNP1-2 are the base of the models proposed. They include a scale driver (Load) and the Percentage of treated load at Bands 1 to 3 and 6. The R^2 of these models is around 0.91. In these models, the Bands are not statistically significant. However, the next set of models (TMSWWWNP3-5) include Pumping Capacity per Main, where the coefficient of this driver is statistically significant and improves the R^2 of the previous two models to around 0.95 and the level of significance of the Bands drivers.

The next three models for the WWWNP, TMSWWWNP6-8 are an extension of model TMSWWWNP3, with different approaches to capturing rainfall drivers. These three approaches all result in statistically significant drivers, increasing the overall performance of the models through the R^2. The results suggest that *urban rainfall* is a strong driver linked to base costs. This might be a reflection that most of the sewerage undertakers, concerned with the collection, treatment, and safe disposal of sewage, comes from 'urban' rainfall that directly influences operational base costs. Reference to total annual rainfall alone could lead to poor representation of the effect of rainfall on assets especially given the significant spatial variation of rainfall across the industry. Overall, the effect of rainfall is relevant for base costs TMSWWWNP as also illustrated in our SWC models, providing consistency across different levels of aggregation.

The last set of models TMSWWWNP9-14 are an extension of models TMSWWWNP4 and TMSWWWNP5. These models improve the R^2 (to around 0.96) compared to the previous ones. All drivers are statistically significant versus the previous models, where some cost drivers were not showing a statistical effect. This result suggests that the inclusion of *urban rainfall* and *pumping capacity per main* to the first two base models TMSWWWNP1-2, significantly improve the overall performance of potential TMSWWWNP models. We believe that models TMSWWWNP9-14 are the more complete candidates to be considered as a new set of WWWNP models.

Retail

For retail we propose 14 models covering both the Bottom-Up and Top-Down models. Of the 14 models proposed, two (see models TMSRDC4 and TMSRTC5) recommend excluding the 2019-20 year from the model and two (see models TMSRDC5 and TMSRTC6) recommend using the period 2013-14 to 2018-19. This could be explained by the external shock to the industry observed in the Transience trend across all companies for year 2019-20, which is likely to be linked to the Covid-19 pandemic. We have noticed that the current PR19 version of the retail models are very unstable with regards to different time periods, and this is apparent when testing the robustness of the models to different sample periods and sizes. We think that the main argument of the retail models centres on *Transience,* which is a material driver for bad debt models. The current PR19 models yield coefficients with signs that appear inconsistent with standard economic logic and with magnitudes lower than expected. We propose either excluding the 2019-20 year from the analysis, using the period 2013-14 to 2017-19 or using a smoothed transience variable by utilising a 3-year moving average of the *Total Migration* driver, for example. Either of these scenarios yield models that are consistent with prior expectations of sign and magnitude of estimated coefficients, which is an improvement from the current PR19 version of the models.

Our second argument focuses on the *Deprivation* variable. In the current PR19 version of the model, this variable performs poorly (i.e. unexpected signs of the coefficients and reduced magnitudes). For this, we propose:

i) Using *Credit Risk Score* to capture deprivation instead of income score unadjusted. This variable performs well in the models and benefits from yearly updated data as opposed to the income score data released every 5 years.

ii) To use income score interpolated in the place of income score unadjusted and

iii) To use a combination of Credit Risk Score and Unemployment Rates to capture deprivation. These scenarios also yield models that are consistent with prior expectations of sign and magnitude of estimated coefficients, which is an improvement from the current PR19 version of the models.

Lastly, we propose that *transience and deprivation* be added as additional drivers to the Other Retail Costs models. This is based on preliminary exploratory modelling of Customer Service Costs, which make up about 50% of the Other Retail Costs in the industry, and for which we argue that it potentially has different drivers from the current PR19 Other Retail Costs, such as *transience and deprivation.* Although these models are promising, they are still a work in progress and not ready to be submitted at this time. However, by adding *transience and deprivation* to the Other Retail Cost model as seen in TMSROC3, the R^2 increases to 0.199 from 0.13 and the dispersion of the efficiency scores is also smaller. This model is also more robust than the PR19 version. Therefore, the absence of *transience and deprivation* from the Other Retail Costs model needs to be re-evaluated based on their performance in the Other Retail Cost models.

We are keen to work with Ofwat to help develop the benchmarking models for PR24 and beyond. I look forward to working with you over the coming months and through the Spring econometric consultation in 2023. We are still working on some Bioresources and Retail models but given the time restriction, we were not able to submit models at this time. However, if we find significant econometric models in the next few weeks, we would request that these models could be considered for the consultation. I attach our detailed response with all the templates, Stata Do files, data and other material requested for this submission. I hope you find these lines useful and if you have any questions or would like to discuss these results further, please contact me at carlos.pinedabermudez@thameswater.co.uk.

Yours sincerely,

Carlos Pineda
Head of Econometric Modelling

Template for submission of econometric models for consultation (Water Resources Plus)

Econometric model formula:

1. TMSWRP1: ln(WRP botex plus Network Reinforcement $_{it}$) = $\alpha + \beta_1$ ln(Properties$_{it}$) + β_2 ln(Capacity_Reservoir_per_Property$_{it}$) + β_3 ln(weighted average density LAD$_{it}$) + β_4 (ln(weighted average density LAD$_{it}$))^2 + β_5 (% Water Treated Complexity 3-6 $_{it}$) + ε_{it}

2. TMSWRP2: ln(WRP botex plus Network Reinforcement $_{it}$) = $\alpha + \beta_1$ ln(Properties$_{it}$) + β_2 ln(Capacity_Reservoir_per_Property$_{it}$) + β_3 ln(weighted average density LAD$_{it}$) + β_4 (ln(weighted average density LAD$_{it}$))^2 + β_5 ln(WAC$_{it}$) + ε_{it}

3. TMSWRP3: ln(WRP botex plus Network Reinforcement $_{it}$) = $\alpha + \beta_1$ ln(Properties$_{it}$) + β_2 ln(weighted average density LAD$_{it}$) + β_3 (ln(weighted average density LAD$_{it}$))^2 + β_4 ln(WACW $_{it}$) + ε_{it}

4. TMSWRP4: ln(WRP botex plus Network Reinforcement $_{it}$) = $\alpha + \beta_1$ ln(Properties$_{it}$) + β_2 ln(Capacity_Reservoir_per_Property$_{it}$) + β_3 ln(weighted average density LAD$_{it}$) + β_4 (ln(weighted average density LAD$_{it}$))^2 + β_5 ln(WACW $_{it}$) + ε_{it}

Description of the dependent variable

All the models use the same definition of **Botex Plus Network Reinforcement** as defined by Ofwat in the Stata code:

Treated Water Distribution

g **botextwd** = BM202TWD + BM336TWD + BM240TWD + BM339ITWD + BM339NITWD + BM339OWD + BC30445TWD + CW00036TWD + W3002TWD + BN4012_TWD - W3032TWD - W3036TWD - APP28RR_W0002 - APP28RR_W0003 - B0201DSWADJ

g **botexplustwd** = **botextwd** + B0201DSITDWNC + B0201DSITDWNO

Wholesale Water

g **botexww** = WS1001CAW + WS01002CAW + WS01004CAW + BM339ICAW_20 + BM339NICAW_20 + BM339OCAW_20 + WS1012CAW + WS1013CAW + W3002CAW_20 + BN4012_WW - W3032TOT - W3036CAW_20 - APP28RR_W0002 - APP28RR_W0003- B0201DSWADJ

g **botexplusww** = **botexww** + B0201DSITDWNC + B0201DSITDWNO

Water Resources Plus

g **botexwrp** = botexww - botextwd

Description of the explanatory variables

- **Ln(Properties) = Natural Log or Ln(Properties) = Ln((BN2221 + BN2161) * 1000);** Number of Properties
- **Ln(WAD_LAD)= Natural Log of Weighted Average Density Local Authority District = Ln(WAD_LAD)=Ln(BN4002):** people per Km2 at LAD level
- **(Ln(WAD_LAD))^2=(Ln(BN4002))^2**
- **Ln(Capacity_Reservoirs_per_Property)= Natural Log of the ratio between:**

$$Ln\ (Capacity\_Reservoir\_per\_Property_{it}) = Ln\left(\frac{Capacity\_Reservoirs_{it}\ \ (Ml)}{Properties_{it}}\right)$$

$$Ln\ (Capacity\_Reservoir\_per\_Property_{it}) = Ln\left(\frac{BN10191}{Properties}\right)$$

- **% proportion of water treated in water treatment works with complexity levels 3-6**
- **watertreated** = CPMW0098 + CPMW0104 + CPMW0110 + CPMW0116 + CPMW0165 + CPMW0166 + CPMW0167 + CPMW0027 + CPMW0033 + CPMW0039 + CPMW0045 + CPMW0185 + CPMW0197 + CPMW0198
- **watertreated36** = CPMW0116 + CPMW0165 + CPMW0166 + CPMW0167 + CPMW0045 + CPMW0185 + CPMW0197 + CPMW0198
- **pctwatertreated36** = (watertreated36 / watertreated) *100
- **wac** = (1*(CPMW0098+CPMW0027)/watertreated) +
  (2*(CPMW0104+CPMW0033)/watertreated) +
  (3*(CPMW0110+CPMW0039)/watertreated) +
  (4*(CPMW0116+CPMW0045)/watertreated) +
  (5*(CPMW0165+CPMW0185)/watertreated) +
  (6*(CPMW0166+CPMW0197)/watertreated) +
  (7*(CPMW0167+CPMW0198)/watertreated)
- **Ln(WAC) = Ln(wac)**
- **WACW= Same as wac but with different weights. We assign a weight of 2 to the simple, band 1 and 2, whereas for complexity bands 3-6 we assign a weight of 5.5. These weights are derived from the simple average of Average(1+2+3)=2 and Average(4+5+6+7)=5.5.**
- **wacw** = (2*(CPMW0098+CPMW0027)/watertreated) +
  (2*(CPMW0104+CPMW0033)/watertreated) +
  (2*(CPMW0110+CPMW0039)/watertreated) +
  (5.5*(CPMW0116+CPMW0045)/watertreated) +
  (5.5*(CPMW0165+CPMW0185)/watertreated) +
  (5.5*(CPMW0166+CPMW0197)/watertreated) +
  (5.5*(CPMW0167+CPMW0198)/watertreated)
- **Ln(WACW)=Ln(wacw)**

Brief comment on the models

- In proposing these models, we have sought to find options to improve the PR19 models' explanatory power and overcome the problem that the PR19 driver representing water treatment complexity is no longer statistically significant. To do this we introduce a new driver "Capacity of Reservoir per Property" and an adjustment into the weights of the wac cost driver as explained in the description of variables section.
- All the models are run using the period of 11 years, 2011-12 to 2021-22.
- All models are more efficient than the PR19 models with lower standard errors providing more confidence on the estimated parameters. Regarding the robustness check tests all models are quite robust to sensitive changes and tests like the RESET specification.
- All models proposed reflect a higher level of R^2 (above 0.91), with all models making significant improvements (TMSWRP1-2 with an R^2 of 0.927 and 0.924, respectively). Furthermore, most of the efficiency scores from the models show a narrower range than observed with the PR19 models.
- Models TMSWRP1-2 include the Capacity of Reservoir (Ml) per Property. This driver captures capital maintenance and operating costs as well as management of reservoirs across water resources. The driver shows a positive effect, and it is statistically significant in the two versions of the models as well as in models TMSWRP3-4. We believe that the Capacity of Reservoir per Property is a significant driver that adds information in explaining base cots, as reservoirs are sensitive to maintain and operate for security reasons (e.g., regular inspection walks, and maintenance is required as stated in the Reservoir Act 1975).
- We note that the Ln(wac) driver is no longer statistically significant in the PR19 model for WRP. To address this, we considered an alternative driver for the effects of water treatment complexity, based on a re-calculation of the weights that are assign to the complexity bands by grouping the lower and upper level of complexity with an average weight. With this adjustment the wac driver shows a statistically significant effect across the specifications in models TMSWRP3-4 alongside with the Capacity of Reservoir (Ml) per Property.

	TMSWRP1	TMSWRP2	TMSWRP3	TMSWRP4
Dependent Variable	Botex + Network Reinforceme nt	Botex + Network Reinforceme nt	Botex + Network Reinforceme nt	Botex + Network Reinforceme nt
Ln(Properties)	1.021*** (0.000)	1.009*** (0.000)	1.074*** (0.000)	1.014*** (0.000)
PCT Water Treated 3-6	0.004*** (0.006)			
Ln(WAD_LAD)	-1.338*** (0.005)	-1.147** (0.013)	-1.579*** (0.000)	-1.282*** (0.007)
(Ln(WAD_LAD))^2	0.087*** (0.005)	0.074** (0.013)	0.099*** (0.001)	0.084*** (0.006)
Ln(Capacity_Reservouir_per_Property)	0.074*** (0.008)	0.085*** (0.004)		0.083*** (0.003)
Ln(WAC)		0.286 (0.262)		
Ln(WACW)			0.534*** (0.003)	0.376** (0.021)
Constant	-5.226*** (0.000)	-5.800*** (0.000)	-5.570*** (0.000)	-5.577*** (0.000)
Estimation Method (OLS or RE)	RE	RE	RE	RE
N (sample size)	187	187	187	187
Model robustness tests				
R2 adjusted	0.927	0.924	0.911	0.925
RESET test	0.532	0.519	0.4	0.548
VIF (max) (OLS)	218.873	201.014	209.338	216.74
Pooling / Chow test (OLS)	0.996	0.998	0.992	0.995
Normality of model residuals (OLS)	0.067	0.052	0.25	0.073
Heteroskedasticity of model residuals (OLS)	0	0	0	0
Test of pooled OLS versus Random Effects (LM test)	0	0	0	0
Efficiency Score Distribution	Min: 0.51 Max: 1.97	Min: 0.48 Max: 1.94	Min: 0.52 Max: 2.01	Min: 0.50 Max: 1.96
Sensitivity of estimated coefficients to removal of most and least efficient company	A	A	G	A

Sensitivity of estimated coefficients to removal of first and last year of the sample	G	A	G	A

Efficiency scores distribution

TMSTWRP1		TMSTWW2		TMSTWW3		TMSTWW4	
SSC	0.51	SSC	0.48	SSC	0.52	SSC	0.50
ANH	0.79	ANH	0.77	PRT	0.69	ANH	0.79
HDD	0.86	PRT	0.86	ANH	0.78	PRT	0.86
PRT	0.87	HDD	0.87	AFW	0.84	HDD	0.86
NES	0.96	NES	0.96	SEW	0.98	NES	0.96
AFW	0.99	TMS	1.01	HDD	1.02	TMS	1.02
TMS	1.01	AFW	1.02	YKY	1.03	AFW	1.02
YKY	1.03	BRL	1.04	TMS	1.06	YKY	1.04
WSH	1.04	WSH	1.04	SVE	1.07	WSH	1.05
BRL	1.06	YKY	1.05	NES	1.08	BRL	1.07
SWB	1.08	SWB	1.08	WSH	1.14	SWB	1.08
SVE	1.08	SVE	1.11	SWB	1.16	SVE	1.09
SEW	1.13	SEW	1.15	NWT	1.18	SEW	1.14
NWT	1.16	NWT	1.16	BRL	1.19	NWT	1.16
WSX	1.36	WSX	1.31	WSX	1.19	WSX	1.32
SES	1.53	SES	1.57	SES	1.71	SES	1.54
SRN	1.97	SRN	1.94	SRN	2.01	SRN	1.96

Comments

- Please indicate the units of the explanatory variable, and whether it was expressed in logs.
- Use asterisks to denote significance level: *** (1%), ** (5%) and * (10%)
- P values should be based on cluster robust standard errors
- In the case of random effects please report Stata's output "R2 overall"
- Please use the following naming convention to assign a name to each model: company acronym, level of aggregation, model number (eg for Anglian Water's wholesale water model number 1: ANHWW1). Please refer to the table below for company acronyms and level of aggregation acronyms.

Company acronyms	Level of aggregation acronyms
Anglian Water: ANH Hafren Dyfrdwy: HDD Northumbrian Water: NES Southern Water: SRN Severn Trent England: SVE South West Water: SWB Thames Water: TMS United Utilities: UUW Dŵr Cymru: WSH Wessex Water: WSX Yorkshire Water: YKY Affinity Water: AFW Bristol Water: BRL Portsmouth Water: PRT SES Water: SES South East Water: SEW South Staffs Water: SSC	**Wholesale water** Treated water distribution: TWD Water resources plus: WRP Water network plus: WWNP Wholesale water: WW **Wholesale wastewater** Sewage collection: SWC Sewage treatment: STW Bioresources: BR Wastewater network plus: WWWNP Bioresources plus: BRP **Residential retail** Bad debt related costs: RDC Other costs: ROC Total costs: RTC

Template for submission of econometric models for consultation (Treated Wated Distribution)

Econometric model formula:

1. TMSTWD1: ln(TWD botex plus Network Reinforcement $_{it}$) = α + β_1 ln(Mains$_{it}$) + β_2 ln(APH_Distribution$_{it}$) + β_3 ln(weighted average density LAD$_{it}$) + β_4 (ln(weighted average density LAD$_{it}$))^2 + ε_{it}

2. TMSTWD2: ln(TWD botex plus Network Reinforcement $_{it}$) = α + β_1 ln(Mains$_{it}$) + β_2 ln(Capacity_Booster_Stn_per_Main$_{it}$) + β_3 ln(weighted average density LAD$_{it}$) + β_4 (ln(weighted average density LAD$_{it}$))^2 + ε_{it}

3. TMSTWD3: ln(TWD botex plus Network Reinforcement $_{it}$) = α + β_1 ln(Mains$_{it}$) + β_2 ln(APH_Distribution$_{it}$) + β_3 ln(weighted average density LAD$_{it}$) + β_4 (ln(weighted average density LAD$_{it}$))^2 + β_5 ln(Weighted_Age_of_Mains$_{it}$) + ε_{it}

4. TMSTWD4: ln(TWD botex plus Network Reinforcement $_{it}$) = α + β_1 ln(Mains$_{it}$) + β_2 ln(Capacity_Booster_Stn_per_Main$_{it}$) + β_3 ln(weighted average density LAD$_{it}$) + β_4 (ln(weighted average density LAD$_{it}$))^2 + β_5 ln(Weighted_Age_of_Mains$_{it}$) + ε_{it}

5. TMSTWD5: ln(TWD botex plus Network Reinforcement $_{it}$) = α + β_1 ln(Mains$_{it}$) + β_2 ln(APH_Distribution$_{it}$) + β_3 ln(weighted average density LAD$_{it}$) + β_4 (ln(weighted average density LAD$_{it}$))^2 + β_5 (% Mains Relined & Renewed$_{it}$) + ε_{it}

6. TMSTWD6: ln(TWD botex plus Network Reinforcement $_{it}$) = α + β_1 ln(Mains$_{it}$) + β_2 ln(Capacity_Booster_Stn_per_Main$_{it}$) + β_3 ln(weighted average density LAD$_{it}$) + β_4 (ln(weighted average density LAD$_{it}$))^2 + β_5 (% Mains Relined & Renewed$_{it}$) + ε_{it}

7. TMSTWD7: ln(TWD botex plus Network Reinforcement $_{it}$) = α + β_1 ln(Mains$_{it}$) + β_2 ln(APH_Distribution$_{it}$) + β_3 ln(weighted average density LAD$_{it}$) + β_4 (ln(weighted average density LAD$_{it}$))^2 + β_5 (% Leakage$_{it}$) + ε_{it}

8. TMSTWD8: ln(TWD botex plus Network Reinforcement $_{it}$) = α + β_1 ln(Mains$_{it}$) + β_2 ln(Capacity_Booster_Stn_per_Main$_{it}$) + β_3 ln(weighted average density LAD$_{it}$) + β_4 (ln(weighted average density LAD$_{it}$))^2 + β_5 (% Leakage$_{it}$) + ε_{it}

9. TMSTWD9: ln(TWD botex plus Network Reinforcement $_{it}$) = α + β_1 ln(Mains$_{it}$) + β_2 ln(APH_Distribution$_{it}$) + β_3 ln(weighted average density LAD$_{it}$) + β_4 (ln(weighted average density LAD$_{it}$))^2 + β_5 (% Mains Relined & Renewed$_{it}$) + β_6 (% Leakage$_{it}$) + ε_{it}

10. TMSTWD10: $\ln(\text{TWD botex plus Network Reinforcement}_{it}) = \alpha + \beta_1 \ln(\text{Mains}_{it}) + \beta_2 \ln(\text{Capacity\_Booster\_Stn\_per\_Main}_{it}) + \beta_3 \ln(\text{weighted average density LAD}_{it}) + \beta_4 (\ln(\text{weighted average density LAD}_{it}))^2 + \beta_5 (\text{\% Mains Relined \& Renewed}_{it}) + \beta_6 (\text{\% Leakage}_{it}) + \varepsilon_{it}$

11. TMSTWD11: $\ln(\text{TWD botex plus Network Reinforcement}_{it}) = \alpha + \beta_1 \ln(\text{Mains}_{it}) + \beta_2 \ln(\text{APH\_Distribution}_{it}) + \beta_3 \ln(\text{weighted average density LAD}_{it}) + \beta_4 (\ln(\text{weighted average density LAD}_{it}))^2 + \beta_5 (\text{\% Mains Relined \& Renewed}_{it}) + \beta_6 \ln(\text{Weighted\_Age\_of\_Mains}_{it}) + \varepsilon_{it}$

12. TMSTWD12: $\ln(\text{TWD botex plus Network Reinforcement}_{it}) = \alpha + \beta_1 \ln(\text{Mains}_{it}) + \beta_2 \ln(\text{Capacity\_Booster\_Stn\_per\_Main}_{it}) + \beta_3 \ln(\text{weighted average density LAD}_{it}) + \beta_4 (\ln(\text{weighted average density LAD}_{it}))^2 + \beta_5 (\text{\% Mains Relined \& Renewed}_{it}) + \beta_6 \ln(\text{Weighted\_Age\_of\_Mains}_{it}) + \varepsilon_{it}$

13. TMSTWD13: $\ln(\text{TWD botex plus Network Reinforcement}_{it}) = \alpha + \beta_1 \ln(\text{Mains}_{it}) + \beta_2 \ln(\text{APH\_Distribution}_{it}) + \beta_3 \ln(\text{weighted average density MSOA\_Population}_{it}) + \beta_4 (\ln(\text{weighted average density MSOA\_Population}_{it}))^2 + \varepsilon_{it}$

14. TMSTWD14: $\ln(\text{TWD botex plus Network Reinforcement}_{it}) = \alpha + \beta_1 \ln(\text{Mains}_{it}) + \beta_2 \ln(\text{Capacity\_Booster\_Stn\_per\_Main}_{it}) + \beta_3 \ln(\text{weighted average density MSOA\_Population}_{it}) + \beta_4 (\ln(\text{weighted average density MSOA\_Population}_{it}))^2 + \varepsilon_{it}$

15. TMSTWD15: $\ln(\text{TWD botex plus Network Reinforcement}_{it}) = \alpha + \beta_1 \ln(\text{Mains}_{it}) + \beta_2 \ln(\text{APH\_Distribution}_{it}) + \beta_3 \ln(\text{weighted average density MSOA\_Population}_{it}) + \beta_4 (\ln(\text{weighted average density MSOA\_Population}_{it}))^2 + \beta_5 (\text{\% Mains Relined \& Renewed}_{it}) + \varepsilon_{it}$

16. TMSTWD16: $\ln(\text{TWD botex plus Network Reinforcement}_{it}) = \alpha + \beta_1 \ln(\text{Mains}_{it}) + \beta_2 \ln(\text{Capacity\_Booster\_Stn\_per\_Main}_{it}) + \beta_3 \ln(\text{weighted average density MSOA\_Population}_{it}) + \beta_4 (\ln(\text{weighted average density MSOA\_Population}_{it}))^2 + \beta_5 (\text{\% Mains Relined \& Renewed}_{it}) + \varepsilon_{it}$

Description of the dependent variable

All the models use the same definition of **Botex Plus Network Reinforcement** as defined by Ofwat in the Stata code:

Treated Water Distribution

g **botextwd** = BM202TWD + BM336TWD + BM240TWD + BM339ITWD + BM339NITWD + BM339OWD + BC30445TWD + CW00036TWD + W3002TWD + BN4012_TWD – W3032TWD – W3036TWD – APP28RR_W0002 – APP28RR_W0003 – B0201DSWADJ

g **botexplustwd** = **botextwd** + B0201DSITDWNC + B0201DSITDWNO

Wholesale Water

g **botexww** = WS1001CAW + WS01002CAW + WS01004CAW + BM339ICAW_20 + BM339NICAW_20 + BM339OCAW_20 + WS1012CAW + WS1013CAW + W3002CAW_20 + BN4012_WW – W3032TOT – W3036CAW_20 – APP28RR_W0002 – APP28RR_W0003– B0201DSWADJ

g **botexplusww** = **botexww** + B0201DSITDWNC + B0201DSITDWNO

Water Resources Plus

g **botexwrp** = botexww – botextwd

Description of the explanatory variables

- **Ln(Length of Mains) = Natural Log or Ln(Mains) = Ln(BN1100);** Km of Mains
- **Ln(Properties) = Natural Log or Ln(Properties) = Ln((BN2221 + BN2161) * 1000);** Number of Properties
- **Ln(APH_Distribution) = Natural Log of Average Pumping Head (Distribution) = Ln(APH_Distribution)=Ln(BN4870);** m.hd.
- **Ln(WAD_LAD)= Natural Log of Weighted Average Density Local Authority District = Ln(WAD_LAD)=Ln(BN4002):** people per Km2 at LAD level
- **(Ln(WAD_LAD))^2=(Ln(BN4002))^2**

- **Ln(Capacity_Booster_Pumping_Stn_per_Main)= Natural Log of the ratio between:**

$$Ln\left(Capacity\_Booster\_Pumping\_Stn\_per\_Main_{it}\right) = Ln\left(\frac{Capacity\_Booster\_Pumping\_Stations_{it}\ (kW)}{Length\ of\ Mains_{it}\ (Km)}\right)$$

$$Ln\,(Capacity\_Booster\_Pumping\_Stn\_per\_Main_{it}) = Ln\left(\frac{BN11300CAP}{BN1100}\right)$$

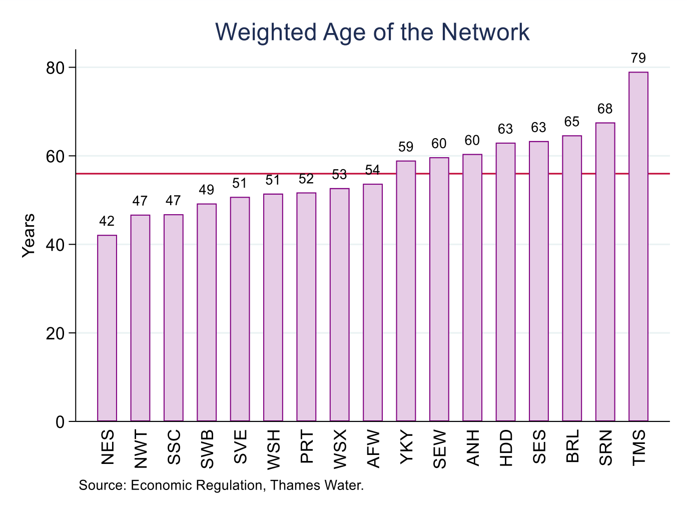
- **Ln(Weighted_Age_of_Mains) = Ln($WAAN_{it}$)** = The following calculation derives the Weighted Average Age of the Network (*WAAN*) for each water company *i* (e.g., TMS, SRN, ANH , etc.) in year *t* (e.g., 2011-12, 2012-13 , …, 2021-22).

$$\begin{aligned} WAAN_{it} = {} & w_{it,1}\left(\frac{MLR_{[Pre\ 1880]}}{Mains_{it}}\right) + w_{it,2}\left(\frac{MLR_{[1881-1900]}}{Mains_{it}}\right) + w_{it,3}\left(\frac{MLR_{[1901-1920]}}{Mains_{it}}\right) \\ & + w_{it,4}\left(\frac{MLR_{[1921-1940]}}{Mains_{it}}\right) + w_{it,5}\left(\frac{MLR_{[1941-1960]}}{Mains_{it}}\right) + w_{it,6}\left(\frac{MLR_{[1961-1980]}}{Mains_{it}}\right) \\ & + w_{it,7}\left(\frac{MLR_{[1981-2000]}}{Mains_{it}}\right) + w_{it,8}\left(\frac{MLR_{[Post\ 2001]}}{Mains_{it}}\right) \end{aligned}$$

Where the weights $w_1, w_{2,\dots}\, w_8$ are defined as the time difference between year *t* and the mid-year within the period of each Mains laid or structurally refurbished (MLR) in Km, BB13000 (pre-1880), BB13010 (1881-1900), BB13020 (1901-1920), BB13030 (1921-1940), BB13040 (1941-1960), BB13050 (1961-1980), BB13060 (1981-2000), BB13070 (Post-2001). For example, for financial year 2014-15 and company *i*, its weight w_2 that correspond to the period [1881-1900] with a mid-year of 1890, is equal to:

$$w_{i(2014-15),2} = 2015 - 1890 = 125\ years$$

These weights multiply the ratio between the Km of mains in each MLR period for a company *i* divided by the Length of Mains (Mains) in Km of that company *i*. The result is a weighted number of years that will reflect how old is the network of each company as illustrated below for the average period 2011-12 to 2021-22:



Source: Economic Regulation, Thames Water.

- **% of Mains Relined & Renewed as proportion of Mains: This is measured as the ratio between:**

$$\% \; of \;\; Mains \; Relined \; \& \; Renewed = \left(\frac{Mains \; Relined \; (Km) + Mains \; Renewed(Km)}{Mains \; (Km)}\right) * 100\%$$
$$= \left(\frac{BN1204 + BN1200}{BN1100}\right) * 100\%$$

- **% Leakage as proportion of Distribution Input: This is measured as the ratio between:**

$$\% \; Leakage \; as \; proportion \; of \; Distribution \; Input = \left(\frac{Leakage \; (Ml/d)}{Distribution \; Input \; (Ml/d)}\right) * 100\%$$
$$= \left(\frac{BN2345A\_CA22\_A}{BN1000\_CA22\_A}\right) * 100\%$$

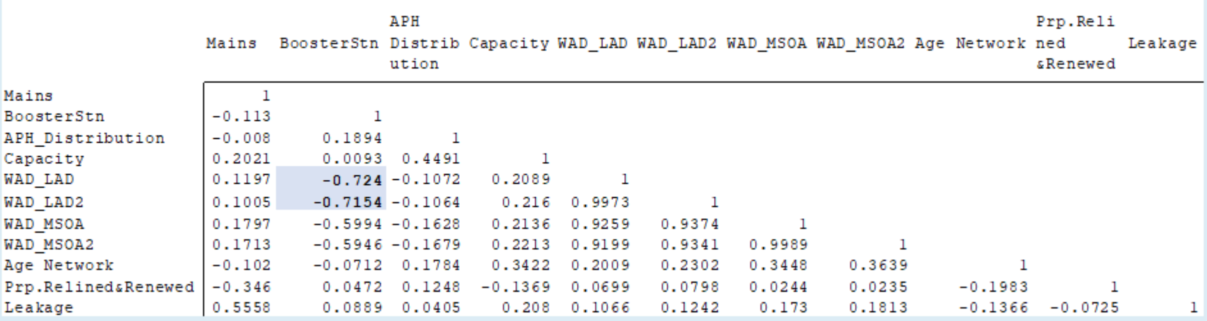
- **Ln(WAD_MSOA_population)= Natural Log of Weighted Average Density at MSOA level = Ln(WAD_LAD)=Ln(BN4000);** people per Km2 at MSOA level
- **(Ln(WAD_MSOA_population))^2=(Ln(BN4000))^2**

Brief comment on the models

- All the models are run using the period of 11 years, 2011-12 to 2021-22.
- All models remain robust (as defined in the guidance) with some marginal but not substantial changes (e.g., sign of coefficient) depending on which dependent variable is chosen:
  - Botex+ (e.g., Opex + Capital Maintenance + Enhancement Growth), used in PR19; or
  - Botex (e.g., Opex + Capital Maintenance).
- All models proposed in TWD improve the R^2 when compared with the current TWD version in PR19.
- Moreover, when using APH or Capacity drivers as substitutes for Number of Booster pumping Stations per main (NBS), the standard errors of the cost drivers in the models proposed, improves significantly relative to the modes that use NBS, providing more confidence in the estimated parameters.

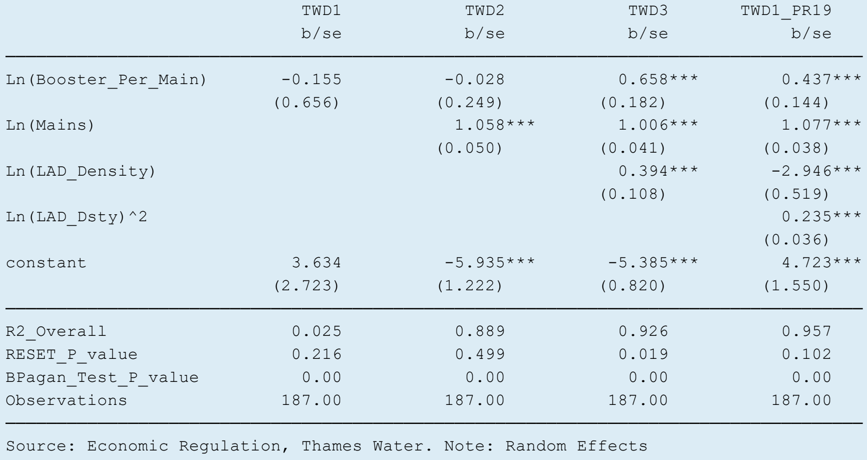
APH or Capacity

- This section explains our concerns on the use of Number of Booster Pumping Stations per Main (NBS) as a driver that explains power costs. We present a brief explanation on the alternatives/substitutes for NBS such as using Average Pumping Head (APH) or Capacity of Pumping Stations (Capacity). We consider that the use of APH or Capacity has a stronger engineering and operational link with base costs than NBS, and that NBS should be replaced by one of these alternatives.
- The following table shows the correlations between the cost drivers used in the set of models (after logarithms or percentages are applied) as defined in the previous section.



	Mains	BoosterStn	APH Distribution	Capacity	WAD_LAD	WAD_LAD2	WAD_MSOA	WAD_MSOA2	Age Network	Prp.Relined &Renewed	Leakage
Mains	1										
BoosterStn	-0.113	1									
APH_Distribution	-0.008	0.1894	1								
Capacity	0.2021	0.0093	0.4491	1							
WAD_LAD	0.1197	**-0.724**	-0.1072	0.2089	1						
WAD_LAD2	0.1005	**-0.7154**	-0.1064	0.216	0.9973	1					
WAD_MSOA	0.1797	-0.5994	-0.1628	0.2136	0.9259	0.9374	1				
WAD_MSOA2	0.1713	-0.5946	-0.1679	0.2213	0.9199	0.9341	0.9989	1			
Age Network	-0.102	-0.0712	0.1784	0.3422	0.2009	0.2302	0.3448	0.3639	1		
Prp.Relined&Renewed	-0.346	0.0472	0.1248	-0.1369	0.0699	0.0798	0.0244	0.0235	-0.1983	1	
Leakage	0.5558	0.0889	0.0405	0.208	0.1066	0.1242	0.173	0.1813	-0.1366	-0.0725	1

- Although a matrix correlation is not the ultimate answer to determine multicollinearity in an econometric model, it is a helpful tool. Among all the drivers we use in the proposed models, there is no indication of extreme degrees of correlation between the drivers presented in the models.
- However, as highlighted in the matrix, there is a reasonably high negative correlation between Booster Pumping Stations per Main (BoosterStn) and density (WAD_LAD) of -0.7. This could indicate that more caution is required when including both WAD_LAD and BoosterStn as explanatory variables.
- Moreover, APH and Capacity that are conceptually related with power costs have a positive correlation of 0.44, whereas the correlation is low for Booster Pumping Stn per Main and APH (0.18) or Capacity (0.009), respectively.
- We explore substitutes for the current driver Number of Booster Pumping Stations per Main (**NBS**) to depict a more coherent link with costs and its engineering and operational rationale.
- We have some concerns relating to the stability of the sign on **NBS** when used in the models and its engineering/economic role as a cost driver. It can be easily explored in the TWD1 model used at PR19 how the sign of **NBS** swap sign when removing the drivers of the model such as mains, or density (see table below). We suspect that this might be related with some degree of multicollinearity between Density and **NBS**. This does not happen when APH or Capacity are used instead in the models.



```
                            TWD1           TWD2           TWD3       TWD1_PR19
                            b/se           b/se           b/se           b/se
-------------------------------------------------------------------------------
Ln(Booster_Per_Main)      -0.155         -0.028          0.658***       0.437***
                         (0.656)        (0.249)        (0.182)        (0.144)
Ln(Mains)                                 1.058***       1.006***       1.077***
                                        (0.050)        (0.041)        (0.038)
Ln(LAD_Density)                                          0.394***      -2.946***
                                                       (0.108)        (0.519)
Ln(LAD_Dsty)^2                                                          0.235***
                                                                      (0.036)
constant                   3.634         -5.935***      -5.385***       4.723***
                         (2.723)        (1.222)        (0.820)        (1.550)
-------------------------------------------------------------------------------
R2_Overall                 0.025          0.889          0.926          0.957
RESET_P_value              0.216          0.499          0.019          0.102
BPagan_Test_P_value         0.00           0.00           0.00           0.00
Observations              187.00         187.00         187.00         187.00
-------------------------------------------------------------------------------
Source: Economic Regulation, Thames Water. Note: Random Effects
```

- The correlation between **botex** and **NBS** is negative. In addition, the correlation with **Density** and **NBS** is also negative (-0.7). These facts could indicate (not claim as this needs to be assessed in a multivariable model) that companies with lower levels of density tend to have more **NBS** (as explored above) and at the same time spend less, whereas companies with high levels of density tend to have less NBS and spend more.
- We think that the strength of **NBS** as a cost driver is not underpinned by a clear engineering-based rationale, and its scope is very limited when used in the models. In particular, NBS is defined in general / aggregate terms and consequently conceals the underpinning factors that are likely to be stronger drivers of cost – in particular the company-specific conditions or the operational characteristics of the stations (e.g., topography).
- What is likely to be a clearer cost driver from an operational, engineering, and economic perspective is the *Capacity of Booster Stations* (as an alternative to APH, in case data is still not robust enough), as one single "Big/Large" Station with large capacity could support or cover more efficiently a high dense area (e.g., to pump water in buildings, and density neighbourhoods etc.) or a spread area. Or it could be more efficient for a company to have more "Small" capacity stations per length of main operating in areas with lower levels of density.
- In other words, the current approach of using **NBS** (instead of APH or Capacity) can be misleading as a proxy to explain power costs, as this is not describing the characteristics of the Stations and how much on average they can reach/cover according to the area where they operate. Moreover, the correlation between Density and *NBS* could be generating multicollinearity issues (e.g., swap in the sign; stability of parameters).
- We believe that *Capacity* might reflect the conditions and needs that each station faces to pump water with different levels of topography etc.
- Finally, using *Capacity* will be aligned with the Capacity driver used in the Wastewater models.
- Finally, the standard errors of the cost drivers and in particular the ones related to APH and Capacity, improve significantly relative to the NBS standard error, providing more confidence in the estimated parameters.
- For the reasons above, we have used either APH or Capacity in place of NBS in each of our proposed models.

Age of the Network, Mains Relined & Renewed and Leakage

- We explore drivers related to capital maintenance (1: age; 2: length of mains relined & renewed) and output-service-cost link (3: leakage).
- It could be argued that these drivers are not entirely exogenous. But at the same time, these drivers are in line with Ofwat's proposal in the Draft and Final Methodology, Appendix 9, p.12 that says: "*we recognise that some drivers are more endogenous than others and are open to considering drivers that are only endogenous in the long term as the risk of perverse incentives is lower*."
- We believe that these three cost drivers have the potential to be consistent with this guidance and should be considered for inclusion.
- The *age of the network* is a relevant cost driver that links to capital maintenance costs. The driver we propose seems to add some level of impact on costs. In the wholesale water aggregate models this shows a statistically significant robust effect.
- *Mains Relined & Renewed* (R&R) could be considered under management control. However, there is a long-term commitment in the industry regarding the rate of replacement in the next decades. This is an important driver for outcomes like leakage or supply interruptions. R&R seems to be statistically significant across all the specifications explored and also under different definitions of the costs (e.g., Botex Plus as in PR19 or Botex).
- *Leakage* is an outcome, or a quality characteristic of the service provided by water companies. Leakage could be considered as a long-term commitment in the industry regarding the reduction by 50% in 2050. Leakage has a clear impact on customers preference (service-cost link) and its impact on costs can be seen in the models proposed. We should expect a negative impact of the leakage rate on botex to reflect the appropriate incentives to reduce leakage. All the estimated effects of the driver yield this negative sign effect.

Coefficient (P-value)	TMSTWD1	TMSTWD2	TMSTWD3	TMSTWD4	TMSTWD5
Dependent Variable	Botex + Network Reinforcement	Botex + Network Reinforcement	Botex + Network Reinforcement	Botex + Network Reinforcement	Botex + Network Reinforcement
Ln(Length of Mains)	1.069*** (0.000)	1.060*** (0.000)	1.071*** (0.000)	1.064*** (0.000)	1.088*** (0.000)
Ln(APH_Distribution)	0.313*** (0.000)		0.278*** (0.000)		0.303*** (0.000)
Ln(WAD_LAD)	-3.203*** (0.000)	-3.021*** (0.000)	-2.713*** (0.000)	-2.646*** (0.000)	-3.379*** (0.000)
(Ln(WAD_LAD))^2	0.245*** (0.000)	0.230*** (0.000)	0.209*** (0.000)	0.202*** (0.000)	0.258*** (0.000)
Ln(Capacity_Booster_Pumping_Stn_per_Main)		0.162*** (0.000)		0.136*** (0.001)	
Ln(Weighted_Age_of_Mains)			0.403* (0.054)	0.34 (0.102)	
% of Mains Relined & Renewed as proportion of Mains					0.157** (0.023)
% Leakage as proportion of DI					
Ln(WAD_MSOA_population)					
(Ln(WAD_MSOA_population))^2					
Constant	2.892* (0.057)	3.728*** (0.003)	-0.193 (0.927)	1.105 (0.582)	3.282** (0.036)
Estimation Method (OLS or RE)	RE	RE	RE	RE	RE
N (sample size)	187	187	187	187	187
Model robustness tests					
R2 adjusted	0.960	0.963	0.964	0.966	0.961
RESET test	0.599	0.144	0.739	0.15	0.258
VIF (max) (OLS)	203	208	243	241	203
Pooling / Chow test (OLS)	0.824	0.964	0.972	0.988	0.622
Normality of model residuals (OLS)	0.918	0.707	0.999	0.772	0.659
Heteroskedasticity of model residuals (OLS)	0.474	0.345	0.329	0.566	0.555

Test of pooled OLS versus Random Effects (LM test)	0.00	0.00	0.00	0.00	0.00
Efficiency Score Distribution	Min: 0.71	Min: 0.75	Min: 0.72	Min: 0.75	Min: 0.75
	Max: 1.33	Max: 1.29	Max: 1.25	Max: 1.23	Max: 1.37
Sensitivity of estimated coefficients to removal of most and least efficient company	G	G	A	A	G
Sensitivity of estimated coefficients to removal of first and last year of the sample	G	G	A	G	A

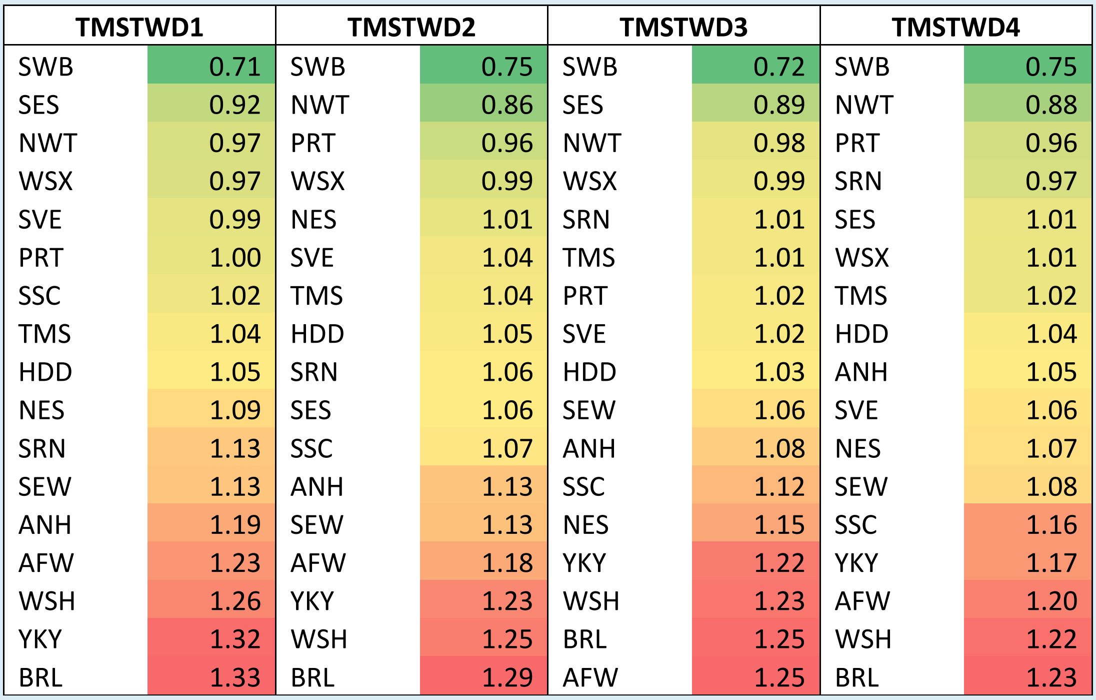
	TMSTWD6	**TMSTWD7**	**TMSTWD8**	**TMSTWD9**	**TMSTWD10**
Dependent Variable	Botex + Network Reinforcement	Botex + Network Reinforcement	Botex + Network Reinforcement	Botex + Network Reinforcement	Botex + Network Reinforcement
Ln(Length of Mains)	1.077*** (0.000)	1.102*** (0.000)	1.087*** (0.000)	1.143*** (0.000)	1.126*** (0.000)
Ln(APH_Distribution)		0.322*** (0.000)		0.316*** (0.000)	
Ln(WAD_LAD)	-3.058*** (0.000)	-3.544*** (0.000)	-3.308*** (0.000)	-3.922*** (0.000)	-3.559*** (0.000)
(Ln(WAD_LAD))^2	0.232*** (0.000)	0.269*** (0.000)	0.251*** (0.000)	0.297*** (0.000)	0.268*** (0.000)
Ln(Capacity_Booster_Pumping_Stn_per_Main)	0.193*** (0.000)		0.157*** (0.000)		0.191*** (0.000)
Ln(Weighted_Age_of_Mains)					
% of Mains Relined & Renewed as proportion of Mains	0.205*** (0.006)			0.192*** (0.009)	0.235*** (0.002)
% Leakage as proportion of DI		-0.013** (0.017)	-0.01 (0.135)	-0.021*** (0.007)	-0.019** (0.024)
Ln(WAD_MSOA_population)					
(Ln(WAD_MSOA_population))^2					
Constant	3.621***	3.987**	4.674***	4.992**	5.241***

	(0.002)	(0.027)	(0.004)	(0.01)	(0.001)
Estimation Method (OLS or RE)	RE	RE	RE	RE	RE
N (sample size)	187	187	187	187	187
Model robustness tests					
R2 adjusted	0.968	0.961	0.964	0.962	0.969
RESET test	0.171	0.827	0.35	0.526	0.401
VIF (max) (OLS)	209	264	268	264	268
Pooling / Chow test (OLS)	0.401	0.961	0.997	0.874	0.706
Normality of model residuals (OLS)	0.178	0.947	0.58	0.776	0.136
Heteroskedasticity of model residuals (OLS)	0.169	0.388	0.288	0.434	0.115
Test of pooled OLS versus Random Effects (LM test)	0.00	0.00	0.00	0.00	0.00
Efficiency Score Distribution	Min: 0.80	Min: 0.71	Min: 0.76	Min: 0.75	Min: 0.80
	Max: 1.30	Max: 1.39	Max: 1.28	Max: 1.49	Max: 1.37
Sensitivity of estimated coefficients to removal of most and least efficient company	A	A	A	A	A
Sensitivity of estimated coefficients to removal of first and last year of the sample	A	A	G	A	A

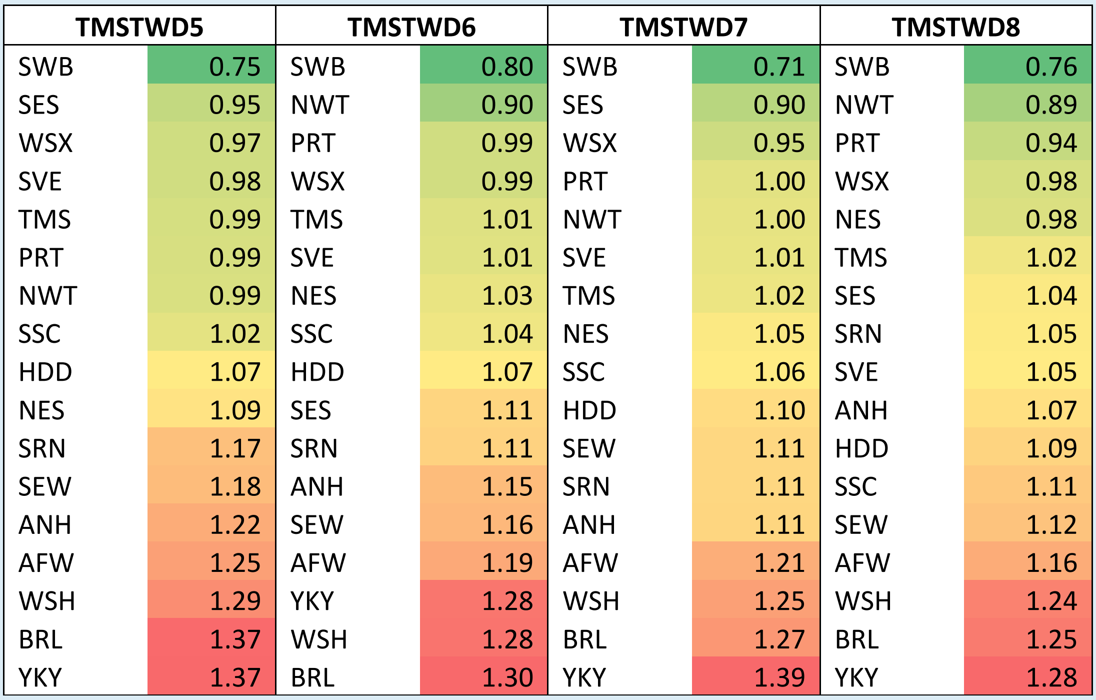
	TMSTWD 11	TMSTWD 12	TMSTWD 13	TMSTWD 14	TMSTWD 15	TMSTWD 16
Dependent Variable	Botex + Network Reinforcement	Botex + Network Reinforcement	Botex + Network Reinforcement	Botex + Network Reinforcement	Botex + Network Reinforcement	Botex + Network Reinforcement
Ln(Length of Mains)	1.097^{***} (0.000)	1.084^{***} (0.000)	1.017^{***} (0.000)	1.012^{***} (0.000)	1.033^{***} (0.000)	1.028^{***} (0.000)
Ln(APH_Distribution)	0.241^{***} (0.007)		0.411^{***} (0.000)		0.397^{***} (0.000)	
Ln(WAD_LAD)	-2.522^{***} (0.000)	-2.271^{***} (0.000)				
(Ln(WAD_LAD))^2	0.194^{***} (0.000)	0.174^{***} (0.000)				

Ln(Capacity_Booster_Pumping_Stn_per_Main)		0.175*** (0.000)		0.144*** (0.001)		0.185*** (0.000)
Ln(Weighted_Age_of_Mains)	0.665* (0.093)	0.568** (0.031)				
% of Mains Relined & Renewed as proportion of Mains	0.232** (0.017)	0.282*** (0.001)			0.178*** (0.008)	0.222*** (0.002)
% Leakage as proportion of DI						
Ln(WAD_MSOA_population)			-6.539*** (0.000)	-5.648*** (0.000)	-6.910*** (0.000)	-5.730*** (0.000)
(Ln(WAD_MSOA_population))^2			0.445*** (0.000)	0.385*** (0.000)	0.470*** (0.000)	0.390*** (0.000)
Constant	-2.004 (0.537)	-1.323 (0.573)	16.573*** (0.000)	15.066*** (0.006)	17.804*** (0.000)	15.124*** (0.003)
Estimation Method (OLS or RE)	RE	RE	RE	RE	RE	RE
N (sample size)	187	187	187	187	187	187
Model robustness tests						
R2 adjusted	0.968	0.973	0.965	0.959	0.967	0.965
RESET test	0.441	0.246	0.719	0.232	0.22	0.32
VIF (max) (OLS)	254	250	492	508	495	511
Pooling / Chow test (OLS)	0.722	0.298	0.767	0.94	0.241	0.076
Normality of model residuals (OLS)	0.899	0.367	0.954	0.631	0.475	0.837
Heteroskedasticity of model residuals (OLS)	0.19	0.23	0.828	0.246	0.79	0.027
Test of pooled OLS versus Random Effects (LM test)	0.00	0.00	0.00	0.00	0.00	0.00
Efficiency Score Distribution	Min: 0.76	Min: 0.80	Min: 0.71	Min: 0.75	Min: 0.75	Min: 0.79
	Max: 1.29	Max: 1.24	Max: 1.32	Max: 1.26	Max: 1.34	Max: 1.26
Sensitivity of estimated coefficients to removal of most and least efficient company	G	G	G	G	A	G
Sensitivity of estimated coefficients to removal of first and last year of the sample	A	A	G	G	A	A

Efficiency scores distribution



TMSTWD1		TMSTWD2		TMSTWD3		TMSTWD4	
SWB	0.71	SWB	0.75	SWB	0.72	SWB	0.75
SES	0.92	NWT	0.86	SES	0.89	NWT	0.88
NWT	0.97	PRT	0.96	NWT	0.98	PRT	0.96
WSX	0.97	WSX	0.99	WSX	0.99	SRN	0.97
SVE	0.99	NES	1.01	SRN	1.01	SES	1.01
PRT	1.00	SVE	1.04	TMS	1.01	WSX	1.01
SSC	1.02	TMS	1.04	PRT	1.02	TMS	1.02
TMS	1.04	HDD	1.05	SVE	1.02	HDD	1.04
HDD	1.05	SRN	1.06	HDD	1.03	ANH	1.05
NES	1.09	SES	1.06	SEW	1.06	SVE	1.06
SRN	1.13	SSC	1.07	ANH	1.08	NES	1.07
SEW	1.13	ANH	1.13	SSC	1.12	SEW	1.08
ANH	1.19	SEW	1.13	NES	1.15	SSC	1.16
AFW	1.23	AFW	1.18	YKY	1.22	YKY	1.17
WSH	1.26	YKY	1.23	WSH	1.23	AFW	1.20
YKY	1.32	WSH	1.25	BRL	1.25	WSH	1.22
BRL	1.33	BRL	1.29	AFW	1.25	BRL	1.23



TMSTWD5		TMSTWD6		TMSTWD7		TMSTWD8	
SWB	0.75	SWB	0.80	SWB	0.71	SWB	0.76
SES	0.95	NWT	0.90	SES	0.90	NWT	0.89
WSX	0.97	PRT	0.99	WSX	0.95	PRT	0.94
SVE	0.98	WSX	0.99	PRT	1.00	WSX	0.98
TMS	0.99	TMS	1.01	NWT	1.00	NES	0.98
PRT	0.99	SVE	1.01	SVE	1.01	TMS	1.02
NWT	0.99	NES	1.03	TMS	1.02	SES	1.04
SSC	1.02	SSC	1.04	NES	1.05	SRN	1.05
HDD	1.07	HDD	1.07	SSC	1.06	SVE	1.05
NES	1.09	SES	1.11	HDD	1.10	ANH	1.07
SRN	1.17	SRN	1.11	SEW	1.11	HDD	1.09
SEW	1.18	ANH	1.15	SRN	1.11	SSC	1.11
ANH	1.22	SEW	1.16	ANH	1.11	SEW	1.12
AFW	1.25	AFW	1.19	AFW	1.21	AFW	1.16
WSH	1.29	YKY	1.28	WSH	1.25	WSH	1.24
BRL	1.37	WSH	1.28	BRL	1.27	BRL	1.25
YKY	1.37	BRL	1.30	YKY	1.39	YKY	1.28

TMSTWD9		TMSTWD10		TMSTWD11		TMSTWD12	
SWB	0.75	SWB	0.80	SWB	0.76	SWB	0.80
SES	0.92	NWT	0.95	SES	0.91	NWT	0.96
WSX	0.94	WSX	0.96	TMS	0.96	SRN	0.98
TMS	0.95	PRT	0.97	SRN	1.00	TMS	1.00
PRT	0.98	TMS	0.98	WSX	1.01	ANH	1.02
SVE	1.00	NES	0.98	PRT	1.02	WSX	1.02
NES	1.03	SVE	1.03	SVE	1.03	PRT	1.03
NWT	1.06	ANH	1.05	NWT	1.04	SES	1.04
SSC	1.08	SES	1.09	ANH	1.05	SVE	1.05
ANH	1.10	SRN	1.10	HDD	1.06	SEW	1.05
SEW	1.15	SSC	1.11	SEW	1.06	HDD	1.06
SRN	1.16	SEW	1.14	SSC	1.19	NES	1.13
HDD	1.16	HDD	1.14	NES	1.20	YKY	1.17
AFW	1.22	AFW	1.17	YKY	1.22	SSC	1.17
WSH	1.27	BRL	1.24	WSH	1.24	BRL	1.20
BRL	1.29	WSH	1.26	BRL	1.25	AFW	1.23
YKY	1.49	YKY	1.37	AFW	1.29	WSH	1.24

TMSTWD13		TMSTWD14		TMSTWD15		TMSTWD16	
SWB	0.71	SWB	0.75	SWB	0.75	SWB	0.79
TMS	0.97	NWT	0.86	TMS	0.92	PRT	0.89
PRT	0.98	PRT	0.86	PRT	0.96	NWT	0.91
SES	0.98	SRN	0.94	SES	1.02	NES	0.98
SRN	1.00	NES	0.96	SVE	1.03	SRN	0.99
NWT	1.01	TMS	1.05	SRN	1.04	TMS	1.00
SEW	1.01	SEW	1.05	NWT	1.05	SVE	1.07
SVE	1.04	SVE	1.10	SEW	1.06	SEW	1.08
NES	1.07	ANH	1.12	NES	1.08	BRL	1.14
SSC	1.10	YKY	1.12	SSC	1.11	ANH	1.15
WSX	1.11	SES	1.14	WSX	1.14	SSC	1.16
BRL	1.13	BRL	1.14	BRL	1.16	YKY	1.16
HDD	1.15	WSX	1.16	HDD	1.19	WSX	1.17
ANH	1.19	HDD	1.20	ANH	1.23	SES	1.19
YKY	1.21	SSC	1.20	YKY	1.25	HDD	1.23
WSH	1.25	WSH	1.22	WSH	1.29	WSH	1.26
AFW	1.32	AFW	1.26	AFW	1.34	AFW	1.26

Comments

- Please indicate the units of the explanatory variable, and whether it was expressed in logs.
- Use asterisks to denote significance level: *** (1%), ** (5%) and * (10%)
- P values should be based on cluster robust standard errors
- In the case of random effects please report Stata's output "R2 overall"
- Please use the following naming convention to assign a name to each model: company acronym, level of aggregation, model number (eg for Anglian Water's wholesale water model number 1: ANHWW1). Please refer to the table below for company acronyms and level of aggregation acronyms.

Company acronyms	Level of aggregation acronyms
Anglian Water: ANH Hafren Dyfrdwy: HDD Northumbrian Water: NES Southern Water: SRN Severn Trent England: SVE South West Water: SWB Thames Water: TMS United Utilities: UUW Dŵr Cymru: WSH Wessex Water: WSX Yorkshire Water: YKY Affinity Water: AFW Bristol Water: BRL Portsmouth Water: PRT SES Water: SES South East Water: SEW South Staffs Water: SSC	**Wholesale water** Treated water distribution: TWD Water resources plus: WRP Water network plus: WWNP Wholesale water: WW **Wholesale wastewater** Sewage collection: SWC Sewage treatment: STW Bioresources: BR Wastewater network plus: WWWNP Bioresources plus: BRP **Residential retail** Bad debt related costs: RDC Other costs: ROC Total costs: RTC

Template for submission of econometric models for consultation (Wholesale Water)

Econometric model formula:

1. TMSWW1: $\ln(\text{WW botex plus Network Reinforcement}_{it}) = \alpha + \beta_1 \ln(\text{Properties}_{it}) + \beta_2 \ln(\text{APH\_Total}_{it}) + \beta_3 \ln(\text{weighted average density LAD}_{it}) + \beta_4 (\ln(\text{weighted average density LAD}_{it}))^2 + \beta_5 (\text{\% Water Treated Complexity 3-6}_{it}) + \varepsilon_{it}$

2. TMSWW2: $\ln(\text{WW botex plus Network Reinforcement}_{it}) = \alpha + \beta_1 \ln(\text{Properties}_{it}) + \beta_2 \ln(\text{APH\_Total}_{it}) + \beta_3 \ln(\text{weighted average density LAD}_{it}) + \beta_4 (\ln(\text{weighted average density LAD}_{it}))^2 + \beta_5 \ln(\text{WAC}_{it}) + \varepsilon_{it}$

3. TMSWW3: $\ln(\text{WW botex plus Network Reinforcement}_{it}) = \alpha + \beta_1 \ln(\text{Properties}_{it}) + \beta_2 \ln(\text{Capacity\_Booster\_Stn\_per\_Main}_{it}) + \beta_3 \ln(\text{weighted average density LAD}_{it}) + \beta_4 (\ln(\text{weighted average density LAD}_{it}))^2 + \beta_5 (\text{\% Water Treated Complexity 3-6}_{it}) + \varepsilon_{it}$

4. TMSWW4: $\ln(\text{WW botex plus Network Reinforcement}_{it}) = \alpha + \beta_1 \ln(\text{Properties}_{it}) + \beta_2 \ln(\text{Capacity\_Booster\_Stn\_per\_Main}_{it}) + \beta_3 \ln(\text{weighted average density LAD}_{it}) + \beta_4 (\ln(\text{weighted average density LAD}_{it}))^2 + \beta_5 \ln(\text{WAC}_{it}) + \varepsilon_{it}$

5. TMSWW5: $\ln(\text{WW botex plus Network Reinforcement}_{it}) = \alpha + \beta_1 \ln(\text{Properties}_{it}) + \beta_2 \ln(\text{Capacity\_Booster\_Stn\_per\_Property}_{it}) + \beta_3 \ln(\text{weighted average density LAD}_{it}) + \beta_4 (\ln(\text{weighted average density LAD}_{it}))^2 + \beta_5 (\text{\% Water Treated Complexity 3-6}_{it}) + \varepsilon_{it}$

6. TMSWW6: $\ln(\text{WW botex plus Network Reinforcement}_{it}) = \alpha + \beta_1 \ln(\text{Properties}_{it}) + \beta_2 \ln(\text{Capacity\_Booster\_Stn\_per\_Property}_{it}) + \beta_3 \ln(\text{weighted average density LAD}_{it}) + \beta_4 (\ln(\text{weighted average density LAD}_{it}))^2 + \beta_5 \ln(\text{WAC}_{it}) + \varepsilon_{it}$

7. TMSWW7: $\ln(\text{WW botex plus Network Reinforcement}_{it}) = \alpha + \beta_1 \ln(\text{Properties}_{it}) + \beta_2 \ln(\text{APH\_Total}_{it}) + \beta_3 \ln(\text{weighted average density LAD}_{it}) + \beta_4 (\ln(\text{weighted average density LAD}_{it}))^2 + \beta_5 (\text{\% Water Treated Complexity 3-6}_{it}) + \beta_6 \ln(\text{Weighted\_Age\_of\_Mains}_{it}) + \varepsilon_{it}$

8. TMSWW8: $\ln(\text{WW botex plus Network Reinforcement}_{it}) = \alpha + \beta_1 \ln(\text{Properties}_{it}) + \beta_2 \ln(\text{APH\_Total}_{it}) + \beta_3 \ln(\text{weighted average density LAD}_{it}) + \beta_4 (\ln(\text{weighted average density LAD}_{it}))^2 + \beta_5 \ln(\text{WAC}_{it}) + \beta_6 \ln(\text{Weighted\_Age\_of\_Mains}_{it}) + \varepsilon_{it}$

9. TMSWW9: ln(WW botex plus Network Reinforcement $_{it}$) = α + β_1 ln(Properties$_{it}$) + β_2 ln(Capacity_Booster_Stn_per_Main$_{it}$) + β_3 ln(weighted average density LAD$_{it}$) + β_4 (ln(weighted average density LAD$_{it}$))^2 + β_5 (% Water Treated Complexity 3-6 $_{it}$) + β_6 ln(Weighted_Age_of_Mains $_{it}$) + ε_{it}

10. TMSWW10: ln(WW botex plus Network Reinforcement $_{it}$) = α + β_1 ln(Properties$_{it}$) + β_2 ln(Capacity_Booster_Stn_per_Main $_{it}$) + β_3 ln(weighted average density LAD$_{it}$) + β_4 (ln(weighted average density LAD$_{it}$))^2 + β_5 ln(WAC$_{it}$) + β_6 ln(Weighted_Age_of_Mains $_{it}$) + ε_{it}

11. TMSWW11: ln(WW botex plus Network Reinforcement $_{it}$) = α + β_1 ln(Properties$_{it}$) + β_2 ln(Capacity_Booster_Stn_per_Property$_{it}$) + β_3 ln(weighted average density LAD$_{it}$) + β_4 (ln(weighted average density LAD$_{it}$))^2 + β_5 (% Water Treated Complexity 3-6 $_{it}$) + β_6 ln(Weighted_Age_of_Mains $_{it}$) + ε_{it}

12. TMSWW12: ln(WW botex plus Network Reinforcement $_{it}$) = α + β_1 ln(Properties$_{it}$) + β_2 ln(Capacity_Booster_Stn_per_Property $_{it}$) + β_3 ln(weighted average density LAD$_{it}$) + β_4 (ln(weighted average density LAD$_{it}$))^2 + β_5 ln(WAC$_{it}$) + β_6 ln(Weighted_Age_of_Mains $_{it}$) + ε_{it}

13. TMSWW13: ln(WW botex plus Network Reinforcement $_{it}$) = α + β_1 ln(CSV$_{it}$) + β_2 ln(APH_Total$_{it}$) + β_3 ln(weighted average density LAD$_{it}$) + β_4 (ln(weighted average density LAD$_{it}$))^2 + β_5 (% Water Treated Complexity 3-6 $_{it}$) + ε_{it}

14. TMSWW14: ln(WW botex plus Network Reinforcement $_{it}$) = α + β_1 ln(CSV$_{it}$) + β_2 ln(APH_Total$_{it}$) + β_3 ln(weighted average density LAD$_{it}$) + β_4 (ln(weighted average density LAD$_{it}$))^2 + β_5 ln(WAC$_{it}$) + ε_{it}

15. TMSWW15: ln(WW botex plus Network Reinforcement $_{it}$) = α + β_1 ln(CSV$_{it}$) + β_2 ln(Capacity_Booster_Stn_per_Main$_{it}$) + β_3 ln(weighted average density LAD$_{it}$) + β_4 (ln(weighted average density LAD$_{it}$))^2 + β_5 (% Water Treated Complexity 3-6 $_{it}$) + ε_{it}

16. TMSWW16: ln(WW botex plus Network Reinforcement $_{it}$) = α + β_1 ln(CSV$_{it}$) + β_2 ln(Capacity_Booster_Stn_per_Main $_{it}$) + β_3 ln(weighted average density LAD$_{it}$) + β_4 (ln(weighted average density LAD$_{it}$))^2 + β_5 ln(WAC$_{it}$) + ε_{it}

17. TMSWW17: ln(WW botex plus Network Reinforcement $_{it}$) = α + β_1 ln(CSV$_{it}$) + β_2 ln(Capacity_Booster_Stn_per_Property$_{it}$) + β_3 ln(weighted average density LAD$_{it}$) + β_4 (ln(weighted average density LAD$_{it}$))^2 + β_5 (% Water Treated Complexity 3-6 $_{it}$) + ε_{it}

18. TMSWW18: ln(WW botex plus Network Reinforcement $_{it}$) = α + β_1 ln(CSV$_{it}$) + β_2 ln(Capacity_Booster_Stn_per_Property $_{it}$) + β_3 ln(weighted average density LAD$_{it}$) + β_4 (ln(weighted average density LAD$_{it}$))^2 + β_5 ln(WAC$_{it}$) + ε_{it}

19. TMSWW19: $\ln(\text{WW botex plus Network Reinforcement}_{it}) = \alpha + \beta_1 \ln(\text{CSV}_{it}) + \beta_2 \ln(\text{APH\_Total}_{it}) + \beta_3 \ln(\text{weighted average density LAD}_{it}) + \beta_4 (\ln(\text{weighted average density LAD}_{it}))^2 + \beta_5 (\text{\% Water Treated Complexity 3–6}_{it}) + \beta_6 \ln(\text{Weighted\_Age\_of\_Mains}_{it}) + \varepsilon_{it}$

20. TMSWW20: $\ln(\text{WW botex plus Network Reinforcement}_{it}) = \alpha + \beta_1 \ln(\text{CSV}_{it}) + \beta_2 \ln(\text{APH\_Total}_{it}) + \beta_3 \ln(\text{weighted average density LAD}_{it}) + \beta_4 (\ln(\text{weighted average density LAD}_{it}))^2 + \beta_5 \ln(\text{WAC}_{it}) + \beta_6 \ln(\text{Weighted\_Age\_of\_Mains}_{it}) + \varepsilon_{it}$

21. TMSWW21: $\ln(\text{WW botex plus Network Reinforcement}_{it}) = \alpha + \beta_1 \ln(\text{CSV}_{it}) + \beta_2 \ln(\text{Capacity\_Booster\_Stn\_per\_Main}_{it}) + \beta_3 \ln(\text{weighted average density LAD}_{it}) + \beta_4 (\ln(\text{weighted average density LAD}_{it}))^2 + \beta_5 (\text{\% Water Treated Complexity 3–6}_{it}) + \beta_6 \ln(\text{Weighted\_Age\_of\_Mains}_{it}) + \varepsilon_{it}$

22. TMSWW22: $\ln(\text{WW botex plus Network Reinforcement}_{it}) = \alpha + \beta_1 \ln(\text{CSV}_{it}) + \beta_2 \ln(\text{Capacity\_Booster\_Stn\_per\_Main}_{it}) + \beta_3 \ln(\text{weighted average density LAD}_{it}) + \beta_4 (\ln(\text{weighted average density LAD}_{it}))^2 + \beta_5 \ln(\text{WAC}_{it}) + \beta_6 \ln(\text{Weighted\_Age\_of\_Mains}_{it}) + \varepsilon_{it}$

23. TMSWW23: $\ln(\text{WW botex plus Network Reinforcement}_{it}) = \alpha + \beta_1 \ln(\text{CSV}_{it}) + \beta_2 \ln(\text{Capacity\_Booster\_Stn\_per\_property}_{it}) + \beta_3 \ln(\text{weighted average density LAD}_{it}) + \beta_4 (\ln(\text{weighted average density LAD}_{it}))^2 + \beta_5 (\text{\% Water Treated Complexity 3–6}_{it}) + \beta_6 \ln(\text{Weighted\_Age\_of\_Mains}_{it}) + \varepsilon_{it}$

24. TMSWW24: $\ln(\text{WW botex plus Network Reinforcement}_{it}) = \alpha + \beta_1 \ln(\text{CSV}_{it}) + \beta_2 \ln(\text{Capacity\_Booster\_Stn\_per\_property}_{it}) + \beta_3 \ln(\text{weighted average density LAD}_{it}) + \beta_4 (\ln(\text{weighted average density LAD}_{it}))^2 + \beta_5 \ln(\text{WAC}_{it}) + \beta_6 \ln(\text{Weighted\_Age\_of\_Mains}_{it}) + \varepsilon_{it}$

25. TMSWW25: $\ln(\text{WW botex plus Network Reinforcement}_{it}) = \alpha + \beta_1 \ln(\text{CSV}_{it}) + \beta_2 \ln(\text{APH\_Total}_{it}) + \beta_3 \ln(\text{weighted average density LAD}_{it}) + \beta_4 (\ln(\text{weighted average density LAD}_{it}))^2 + \beta_5 \ln(\text{WACW}_{it}) + \beta_6 \ln(\text{Weighted\_Age\_of\_Mains}_{it}) + \varepsilon_{it}$

26. TMSWW26: $\ln(\text{WW botex plus Network Reinforcement}_{it}) = \alpha + \beta_1 \ln(\text{CSV}_{it}) + \beta_2 \ln(\text{Capacity\_Booster\_Stn\_per\_main}_{it}) + \beta_3 \ln(\text{weighted average density LAD}_{it}) + \beta_4 (\ln(\text{weighted average density LAD}_{it}))^2 + \beta_5 \ln(\text{WACW}_{it}) + \beta_6 \ln(\text{Weighted\_Age\_of\_Mains}_{it}) + \varepsilon_{it}$

27. TMSWW27: $\ln(\text{WW botex plus Network Reinforcement}_{it}) = \alpha + \beta_1 \ln(\text{CSV}_{it}) + \beta_2 \ln(\text{Capacity\_Booster\_Stn\_per\_property}_{it}) + \beta_3 \ln(\text{weighted average density LAD}_{it}) + \beta_4 (\ln(\text{weighted average density LAD}_{it}))^2 + \beta_5 \ln(\text{WACW}_{it}) + \beta_6 \ln(\text{Weighted\_Age\_of\_Mains}_{it}) + \varepsilon_{it}$

Description of the dependent variable

All the models use the same definition of **Botex Plus Network Reinforcement** as defined by Ofwat in the Stata code:

Treated Water Distribution

g **botextwd** = BM202TWD + BM336TWD + BM240TWD + BM339ITWD + BM339NITWD + BM339OWD + BC30445TWD + CW00036TWD + W3002TWD + BN4012_TWD – W3032TWD – W3036TWD – APP28RR_W0002 – APP28RR_W0003 – B0201DSWADJ

g **botexplustwd** = **botextwd** + B0201DSITDWNC + B0201DSITDWNO

Wholesale Water

g **botexww** = WS1001CAW + WS01002CAW + WS01004CAW + BM339ICAW_20 + BM339NICAW_20 + BM339OCAW_20 + WS1012CAW + WS1013CAW + W3002CAW_20 + BN4012_WW – W3032TOT – W3036CAW_20 – APP28RR_W0002 – APP28RR_W0003– B0201DSWADJ

g **botexplusww** = **botexww** + B0201DSITDWNC + B0201DSITDWNO

Water Resources Plus

g **botexwrp** = botexww – botextwd

Description of the explanatory variables

- **Ln(Length of Mains) = Natural Log or Ln(Mains) = Ln(BN1100);** Km of Mains
- **Ln(Properties) = Natural Log or Ln(Properties) = Ln((BN2221 + BN2161) * 1000);** Number of Properties
- **APH_total = BN4861 + BN4862 + BN10902 +BN4870**
- **Ln(APH_total) = Natural Log of Average Pumping Head (Total) = Ln(APH_total)=Ln(BN4861 + BN4862 + BN10902 +BN4870);** m.hd.
- **Ln(WAD_LAD)= Natural Log of Weighted Average Density Local Authority District = Ln(WAD_LAD)=Ln(BN4002):** people per Km2 at LAD level
- **(Ln(WAD_LAD))^2=(Ln(BN4002))^2**

- **Ln(Capacity_Booster_Pumping_Stn_per_Main)= Natural Log of the ratio between:**

$$Ln\ (Capacity\_Booster\_Pumping\_Stn\_per\_Main_{it}) = Ln\left(\frac{Capacity\_Booster\_Pumping\_Stations_{it}\ (kW)}{Length\ of\ Mains_{it}\ (Km)}\right)$$

$$Ln\ (Capacity\_Booster\_Pumping\_Stn\_per\_Main_{it}) = Ln\left(\frac{BN11300CAP}{BN1100}\right)$$

- **Ln(Capacity_Booster_Pumping_Stn_per_Property)= Natural Log of the ratio between:**

$$Ln\ (Capacity\_Booster\_Pumping\_Stn\_per\_Property_{it}) = Ln\left(\frac{Capacity\_Booster\_Pumping\_Stations_{it}\ (kW)}{Properties_{it}}\right)$$

$$Ln\ (Capacity\_Booster\_Pumping\_Stn\_per\_Property_{it}) = Ln\left(\frac{BN11300CAP}{Properties}\right)$$

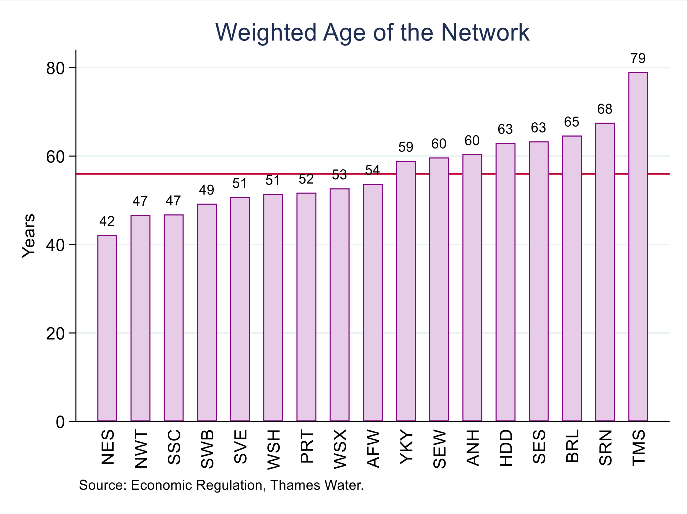
- **Ln(Weighted_Age_of_Mains) = Ln($\boldsymbol{WAAN_{it}}$) =** The following calculation derives the Weighted Average Age of the Network (*WAAN*) for each water company *i* (e.g., TMS, SRN, ANH , etc.) in year *t* (e.g., 2011-12, 2012-13 , …, 2021-22).

$$WAAN_{it} = w_{it,1}\left(\frac{MLR_{[Pre\ 1880]}}{Mains_{it}}\right) + w_{it,2}\left(\frac{MLR_{[1881-1900]}}{Mains_{it}}\right) + w_{it,3}\left(\frac{MLR_{[1901-1920]}}{Mains_{it}}\right) + w_{it,4}\left(\frac{MLR_{[1921-1940]}}{Mains_{it}}\right) + w_{it,5}\left(\frac{MLR_{[1941-1960]}}{Mains_{it}}\right) + w_{it,6}\left(\frac{MLR_{[1961-1980]}}{Mains_{it}}\right) + w_{it,7}\left(\frac{MLR_{[1981-2000]}}{Mains_{it}}\right) + w_{it,8}\left(\frac{MLR_{[Post\ 2001]}}{Mains_{it}}\right)$$

Where the weights $w_1, w_{2,\dots}\ w_8$ are defined as the distance between year *t* and the mid-year within the period of each *Mains laid or structurally refurbished* ($\boldsymbol{MLR}$) in Km, BB13000 (pre-1880), BB13010 (1881-1900), BB13020 (1901-1920), BB13030 (1921-1940), BB13040 (1941-1960), BB13050 (1961-1980), BB13060 (1981-2000), BB13070 (Post-2001). For example, for financial year 2014-15 and company *i*, its weight w_2 that correspond to the period [1881-1900] with a mid-year of 1890, is equal to:

$$w_{i(2014-15),2} = 2015 - 1890 = 125\ years$$

All companies will have the same weight for any range. These weights multiply the ratio between the Km of mains in each MLR period for a company *i* divided by the Length of Mains (Mains) in Km of that company *i*. The result is a weighted number of years that will reflect how old is the network of each company as illustrated in the average for the period 2011-12 to 2021-22 below:



Source: Economic Regulation, Thames Water.

- **Proportion_Natural_Water_Resources= prp_river_abst + prp_boreholes + prp_acquifer: or BN4838 + BN4848 + BN4847**
- **% proportion of water treated in water treatment works with complexity levels 3-6**
- **watertreated** = CPMW0098 + CPMW0104 + CPMW0110 + CPMW0116 + CPMW0165 + CPMW0166 + CPMW0167 + CPMW0027 + CPMW0033 + CPMW0039 + CPMW0045 + CPMW0185 + CPMW0197 + CPMW0198
- **watertreated36** = CPMW0116 + CPMW0165 + CPMW0166 + CPMW0167 + CPMW0045 + CPMW0185 + CPMW0197 + CPMW0198
- **pctwatertreated36** = (watertreated36 / watertreated) *100
- **wac** = (1*(CPMW0098+CPMW0027)/watertreated) +
(2*(CPMW0104+CPMW0033)/watertreated) +
(3*(CPMW0110+CPMW0039)/watertreated) +
(4*(CPMW0116+CPMW0045)/watertreated) +
(5*(CPMW0165+CPMW0185)/watertreated) +
(6*(CPMW0166+CPMW0197)/watertreated) +
(7*(CPMW0167+CPMW0198)/watertreated)
- **Ln(WAC)=Ln(wac)**
- **Composite Scale Variable (for Wholesale Water) (CSV) =** (properties)^0.5 * (Mains)^0.5; We allocate the same weight to mains and properties, although these weights are flexible to other figures.
- **Ln(CSV)= Natural Log or Ln(CSV)**
- **WACW= Same as wac but with different weights. We assign a weight of 2 to the simple, band 1 and 2, whereas for complexity bands 3-6 we assign a weight of 5.5. These weights are derived from the simple average of Average(1+2+3)=2 and Average(4+5+6+7)=5.5.**
- **wacw** = (2*(CPMW0098+CPMW0027)/watertreated) +
(2*(CPMW0104+CPMW0033)/watertreated) +
(2*(CPMW0110+CPMW0039)/watertreated) +
(5.5*(CPMW0116+CPMW0045)/watertreated) +
(5.5*(CPMW0165+CPMW0185)/watertreated) +

(5.5*(CPMW0166+CPMW0197)/watertreated) + (5.5*(CPMW0167+CPMW0198)/watertreated)

- **Ln(WACW)=Ln(wacw)**

Brief comment on the models

- The aim of these section is to show how the WW models can be improved when compared to the PR19 models using some of the insights generated from the disaggregated models WRP and TWD. We believe that the use of APH or Capacity as substitutes of Booster Pumping Station per Main is feasible and will provide significant improvements to the WW models. Moreover, a CSV cost driver could be an alternative scale driver for the aggregate WW models, where results seem to suggest improvements in the R^2.
- All the models are run using the period of 11 years, 2011-12 to 2021-22.
- All models remain robust (as defined in the guidance) with some marginal but not substantial changes (e.g.no change in the sign of a coefficient) when the dependent variable is either the one used at PR19: Botex + (e.g., Opex + Capital Maintenance + Enhancement Growth) or Botex (e.g., Opex + Capital Maintenance). Models are generally more robust when using the Botex+ PR19 definition.
- About 21 of our proposed models in WW improve or maintain the R^2 when compared with the current WW version models at PR19. The rest of the models remain with an R^2 greater than 0.967 and they are important to explain the narrative of the models and new drivers proposed. Furthermore, the efficiency scores of several models proposed show less variability with respect to the PR19 models.
- We proposed alternative cost drivers in the TWD models to replace the *Number of Booster Pumping Station per Main* **(NBS).** To maintain consistency across the sets of models we have modified the current set of WW PR19 models using either Total Average Pumping Head (APH), Capacity of Booster Pumping Station per Main or per Property as substitutes for NBS.
- Models TMSWW 1-6 represent these changes. In general, these models perform well statistically. They also show a significant effect for any of the substitutes of NBS, and estimated coefficients have the expected sign. Furthermore, the proposed models are consistent with the Capacity measure used in the wholesale wastewater models and Treated Water Distribution (TWD).
- We extend models TMSWW1 to 6 with models TWSWW7-12 adding the *age of the network* as we do in the TWD models. These extended models improve the R^2 when compared to the PR19 models, with only marginal changes in some robustness checks (e.g., removing the last and first year of the sample or the most/less efficient company).

- We also introduce for models TMSWW13-27 the same set of substitute drivers for NBS and a substitute for the scale driver Properties. We introduce a Composite Scale Variable (CSV) cost driver. The combination of the CSV and APH, Capacity and Age of the Network improve the R^2 versus the current PR19 models. Moreover, the efficiency of the models is also improved (lower standard errors).
- All models that use CSV also improve the level of significance of all the cost drivers used in the models when compared to the models that use Properties (as in the PR19 models) as the scale driver (TMSWW1-12).
- Regarding the weights on the CSV we allocate 0.50 for each of the components of the CSV, Length of Mains and Number of Properties. Clearly these weights are flexible and can be changed. The idea is to introduce another dimension of the output in the scale drivers that are highly correlated into a single variable. It seems that the introduction of this CSV driver improves the performance of the models at the aggregate level of the wholesale water base costs.
- Finally, models TMSWW25-27 introduce the adjusted WAC that we present in Water Resources Plus. This is consistent with our approach taken in the water resources plus models, which also shows improvements in explanatory power with an R^2 of 0.974.

	TMSWW1	TMSWW2	TMSWW3	TMSWW4
Dependent Variable	Botex + Network Reinforcement	Botex + Network Reinforcement	Botex + Network Reinforcement	Botex + Network Reinforcement
Ln(Properties)	1.096*** 0	1.087*** 0	1.062*** 0	1.053*** 0
% Water Treated complexity 3-6	0.003* 0.059		0.003** 0.013	
Ln(APH_Total)	0.323*** 0.008	0.309*** 0.008		
Ln(WAD_LAD)	-2.579*** 0	-2.401*** 0	-2.271*** 0	-2.058*** 0
(Ln(WAD_LAD))^2	0.177*** 0	0.164*** 0	0.153*** 0	0.137*** 0
Ln(WAC)		0.280*		0.336**

		0.094		0.038
Ln(Capacity_Booster_Pumping_Stn_per_Main)			0.097** 0.031	0.097** 0.046
Ln(Capacity_Booster_Pumping_Stn_per_Property)				
Ln(Weighted_Age_of_Mains)				
Ln(CSV)				
Ln(WACW)				
Constant	-2.888* 0.072	-3.486** 0.021	-1.876* 0.088	-2.695** 0.01
Estimation Method (OLS or RE)	RE	RE	RE	RE
N (sample size)	187	187	187	187
Model robustness tests				
R2 adjusted	0.969	0.969	0.967	0.967
RESET test	0.786	0.824	0.419	0.4
VIF (max) (OLS)	210.907	205.345	223.698	207.428
Pooling / Chow test (OLS)	0.592	0.534	0.931	0.762
Normality of model residuals (OLS)	0.069	0.439	0.067	0.577
Heteroskedasticity of model residuals (OLS)	0	0	0	0
Test of pooled OLS versus Random Effects (LM test)	0	0	0	0
Efficiency Score Distribution	Min: 0.81	Min: 0.78	Min: 0.83	Min: 0.80
	Max: 1.33	Max: 1.32	Max: 1.41	Max: 1.39
Sensitivity of estimated coefficients to removal of most and least efficient company	A	A	A	A
Sensitivity of estimated coefficients to removal of first and last year of the sample	A	A	A	A

	TMSWW5	TMSWW6	TMSWW7	TMSWW8
Dependent Variable	Botex + Network Reinforcement	Botex + Network Reinforcement	Botex + Network Reinforcement	Botex + Network Reinforcement
Ln(Properties)	1.059*** 0	1.051*** 0	1.092*** 0	1.087*** 0
% Water Treated complexity 3-6	0.003** 0.017		0.003* 0.085	
Ln(APH_Total)			0.251** 0.023	0.252** 0.021
Ln(WAD_LAD)	-2.277*** 0	-2.078*** 0	-2.043*** 0	-1.928*** 0
(Ln(WAD_LAD))^2	0.155*** 0	0.141*** 0	0.137*** 0	0.129*** 0
Ln(WAC)		0.327** 0.046		0.228 0.14
Ln(Capacity_Booster_Pumping_Stn_per_Main)				
Ln(Capacity_Booster_Pumping_Stn_per_Property)	0.106** 0.023	0.108** 0.027		
Ln(Weighted_Age_of_Mains)			0.390*** 0.001	0.372*** 0.001
Ln(CSV)				
Ln(WACW)				
Constant	-1.477 0.127	-2.240** 0.02	-5.774*** 0.002	-6.154*** 0
Estimation Method (OLS or RE)	RE	RE	RE	RE
N (sample size)	187	187	187	187
Model robustness tests				
R2 adjusted	0.968	0.968	0.972	0.972
RESET test	0.404	0.372	0.692	0.76
VIF (max) (OLS)	220.064	204.069	273.993	262.66
Pooling / Chow test (OLS)	0.953	0.842	0.79	0.704

Normality of model residuals (OLS)	0.088	0.659	0.08	0.353
Heteroskedasticity of model residuals (OLS)	0	0	0	0
Test of pooled OLS versus Random Effects (LM test)	0	0	0	0
Efficiency Score Distribution	Min: 0.83	Min: 0.80	Min: 0.88	Min: 0.86
	Max: 1.42	Max: 1.39	Max: 1.22	Max: 1.21
Sensitivity of estimated coefficients to removal of most and least efficient company	A	A	G	G
Sensitivity of estimated coefficients to removal of first and last year of the sample	G	A	A	A

	TMSWW9	**TMSWW10**	**TMSWW11**	**TMSWW12**
Dependent Variable	Botex + Network Reinforcement	Botex + Network Reinforcement	Botex + Network Reinforcement	Botex + Network Reinforcement
Ln(Properties)	1.067*** 0	1.059*** 0	1.064*** 0	1.057*** 0
% Water Treated complexity 3-6	0.003** 0.016		0.003** 0.022	
Ln(APH_Total)				
Ln(WAD_LAD)	-1.776*** 0	-1.629*** 0	-1.785*** 0	-1.651*** 0
(Ln(WAD_LAD))^2	0.116*** 0	0.105*** 0	0.118*** 0	0.109*** 0
Ln(WAC)		0.282** 0.044		0.271* 0.055
Ln(Capacity_Booster_Pumping_Stn_per_Main)	0.065 0.118	0.068 0.137		
Ln(Capacity_Booster_Pumping_Stn_per_Property)			0.076* 0.077	0.081* 0.078
Ln(Weighted_Age_of_Mains)	0.417***	0.391***	0.409***	0.381***

	0.006	0.002	0.005	0.003
Ln(CSV)				
Ln(WACW)				
Constant	-5.197*** 0.002	-5.664*** 0	-4.856*** 0.003	-5.247*** 0
Estimation Method (OLS or RE)	RE	RE	RE	RE
N (sample size)	187	187	187	187
Model robustness tests				
R2 adjusted	0.972	0.971	0.972	0.972
RESET test	0.56	0.661	0.532	0.586
VIF (max) (OLS)	266.681	246.118	263.161	243.521
Pooling / Chow test (OLS)	0.927	0.759	0.945	0.824
Normality of model residuals (OLS)	0.119	0.519	0.176	0.619
Heteroskedasticity of model residuals (OLS)	0	0	0	0.001
Test of pooled OLS versus Random Effects (LM test)	0	0	0	0
Efficiency Score Distribution	Min: 0.83	Min: 0.82	Min: 0.84	Min: 0.83
	Max: 1.27	Max: 1.26	Max: 1.28	Max: 1.26
Sensitivity of estimated coefficients to removal of most and least efficient company	A	A	A	A
Sensitivity of estimated coefficients to removal of first and last year of the sample	A	A	A	A

	TMSWW13	TMSWW14	TMSWW15	TMSWW16
Dependent Variable	Botex + Network Reinforcement	Botex + Network Reinforcement	Botex + Network Reinforcement	Botex + Network Reinforcement
Ln(Properties)				
% Water Treated complexity 3-6	0.003** 0.04		0.003*** 0.01	
Ln(APH_Total)	0.310*** 0.005	0.291*** 0.005		
Ln(WAD_LAD)	-2.698*** 0	-2.519*** 0	-2.357*** 0	-2.166*** 0
(Ln(WAD_LAD))^2	0.196*** 0	0.183*** 0	0.169*** 0	0.155*** 0
Ln(WAC)		0.293* 0.068		0.336** 0.031
Ln(Capacity_Booster_Pumping_Stn_per_Main)			0.108** 0.014	0.106** 0.024
Ln(Capacity_Booster_Pumping_Stn_per_Property)				
Ln(Weighted_Age_of_Mains)				
Ln(CSV)	1.088*** 0	1.079*** 0	1.053*** 0	1.045*** 0
Ln(WACW)				
Constant	-0.527 0.654	-1.142 0.294	0.255 0.787	-0.526 0.546
Estimation Method (OLS or RE)	RE	RE	RE	RE
N (sample size)	187	187	187	187
Model robustness tests				
R2 adjusted	0.972	0.972	0.971	0.971
RESET test	0.674	0.722	0.364	0.334
VIF (max) (OLS)	212.029	207.408	224.57	208.601
Pooling / Chow test (OLS)	0.678	0.615	0.935	0.856

Normality of model residuals (OLS)	0.188	0.625	0.042	0.295
Heteroskedasticity of model residuals (OLS)	0	0	0	0
Test of pooled OLS versus Random Effects (LM test)	0	0	0	0
Efficiency Score Distribution	Min: 0.82	Min: 0.80	Min: 0.84	Min: 0.81
	Max: 1.26	Max: 1.34	Max: 1.43	Max: 1.41
Sensitivity of estimated coefficients to removal of most and least efficient company	A	A	A	A
Sensitivity of estimated coefficients to removal of first and last year of the sample	A	A	A	A

	TMSWW17	TMSWW18	TMSWW19	TMSWW20
Dependent Variable	Botex + Network Reinforcement	Botex + Network Reinforcement	Botex + Network Reinforcement	Botex + Network Reinforcement
Ln(Properties)				
% Water Treated complexity 3-6	0.003** 0.012		0.003* 0.057	
Ln(APH_Total)			0.235** 0.013	0.232** 0.014
Ln(WAD_LAD)	-2.380*** 0	-2.192*** 0	-2.158*** 0	-2.043*** 0
(Ln(WAD_LAD))^2	0.173*** 0	0.159*** 0	0.156*** 0	0.147*** 0
Ln(WAC)		0.331** 0.036		0.240* 0.1
Ln(Capacity_Booster_Pumping_Stn_per_Main)				

Ln(Capacity_Booster_Pumping_Stn_per_Property)	0.110** 0.015	0.109** 0.023		
Ln(Weighted_Age_of_Mains)			0.397*** 0	0.375*** 0
Ln(CSV)	1.051*** 0	1.043*** 0	1.084*** 0	1.078*** 0
Ln(WACW)				
Constant	0.718 0.42	-0.052 0.953	-3.445** 0.023	-3.818*** 0.004
Estimation Method (OLS or RE)	RE	RE	RE	RE
N (sample size)	187	187	187	187
Model robustness tests				
R2 adjusted	0.971	0.971	0.975	0.975
RESET test	0.346	0.318	0.565	0.61
VIF (max) (OLS)	221.187	205.436	276.151	265.525
Pooling / Chow test (OLS)	0.945	0.885	0.779	0.673
Normality of model residuals (OLS)	0.041	0.257	0.799	0.905
Heteroskedasticity of model residuals (OLS)	0	0	0.005	0.011
Test of pooled OLS versus Random Effects (LM test)	0	0	0	0
Efficiency Score Distribution	Min: 0.84 Max: 1.43	Min: 0.81 Max: 1.41	Min: 0.87 Max: 1.24	Min: 0.85 Max: 1.23
Sensitivity of estimated coefficients to removal of most and least efficient company	A	A	A	A
Sensitivity of estimated coefficients to removal of first and last year of the sample	A	A	G	A

	TMSWW21	TMSWW22	TMSWW23	TMSWW24
Dependent Variable	Botex + Network Reinforcement	Botex + Network Reinforcement	Botex + Network Reinforcement	Botex + Network Reinforcement
Ln(Properties)				
% Water Treated complexity 3-6	0.003** 0.011		0.003** 0.014	
Ln(APH_Total)				
Ln(WAD_LAD)	-1.878*** 0	-1.749*** 0	-1.895*** 0	-1.770*** 0
(Ln(WAD_LAD))^2	0.134*** 0	0.124*** 0	0.136*** 0	0.127*** 0
Ln(WAC)		0.279** 0.034		0.274** 0.04
Ln(Capacity_Booster_Pumping_Stn_per_Main)	0.076** 0.048	0.078* 0.072		
Ln(Capacity_Booster_Pumping_Stn_per_Property)			0.078* 0.051	0.081* 0.069
Ln(Weighted_Age_of_Mains)	0.417*** 0.002	0.389*** 0.001	0.416*** 0.002	0.387*** 0.001
Ln(CSV)	1.059*** 0	1.052*** 0	1.057*** 0	1.050*** 0
Ln(WACW)				
Constant	-3.002** 0.043	-3.432*** 0.009	-2.665* 0.069	-3.067** 0.022
Estimation Method (OLS or RE)	RE	RE	RE	RE
N (sample size)	187	187	187	187
Model robustness tests				
R2 adjusted	0.974	0.974	0.975	0.974
RESET test	0.485	0.491	0.471	0.492
VIF (max) (OLS)	268.152	247.647	264.548	245.044
Pooling / Chow test (OLS)	0.901	0.783	0.907	0.8

Normality of model residuals (OLS)	0.587	0.987	0.626	0.969
Heteroskedasticity of model residuals (OLS)	0.004	0.008	0.005	0.01
Test of pooled OLS versus Random Effects (LM test)	0	0	0	0
Efficiency Score Distribution	Min: 0.86	Min: 0.85	Min: 0.86	Min: 0.85
	Max: 1.29	Max: 1.28	Max: 1.29	Max: 1.28
Sensitivity of estimated coefficients to removal of most and least efficient company	A	A	A	A
Sensitivity of estimated coefficients to removal of first and last year of the sample	A	A	A	A

	TMSWW25	TMSWW26	TMSWW27
Dependent Variable	Botex + Network Reinforcement	Botex + Network Reinforcement	Botex + Network Reinforcement
Ln(Properties)			
% Water Treated complexity 3-6			
Ln(APH_Total)	0.247** 0.014		
Ln(WAD_LAD)	-2.129*** 0	-1.833*** 0	-1.851*** 0
(Ln(WAD_LAD))^2	0.154*** 0	0.130*** 0	0.133*** 0
Ln(WAC)			
Ln(Capacity_Booster_Pumping_Stn_per_Main)		0.080* 0.054	
Ln(Capacity_Booster_Pumping_Stn_per_Property)			0.083* 0.054
Ln(Weighted_Age_of_Mains)	0.407*** 0	0.427*** 0.001	0.426*** 0.001

Ln(CSV)	1.085*** 0	1.058*** 0	1.056*** 0
Ln(WACW)	0.234 0.145	0.282** 0.041	0.274** 0.049
Constant	-3.810*** 0.009	-3.390** 0.018	-3.021** 0.034
Estimation Method (OLS or RE)	RE	RE	RE
N (sample size)	187	187	187
Model robustness tests			
R2 adjusted	0.974	0.974	0.974
RESET test	0.577	0.465	0.448
VIF (max) (OLS)	274.654	267.337	263.48
Pooling / Chow test (OLS)	0.76	0.885	0.894
Normality of model residuals (OLS)	0.869	0.687	0.729
Heteroskedasticity of model residuals (OLS)	0.004	0.003	0.004
Test of pooled OLS versus Random Effects (LM test)	0	0	0
Efficiency Score Distribution	Min: 0.86 Max: 1.23	Min: 0.84 Max: 1.28	Min: 0.85 Max: 1.28
Sensitivity of estimated coefficients to removal of most and least efficient company	A	A	A
Sensitivity of estimated coefficients to removal of first and last year of the sample	A	A	A

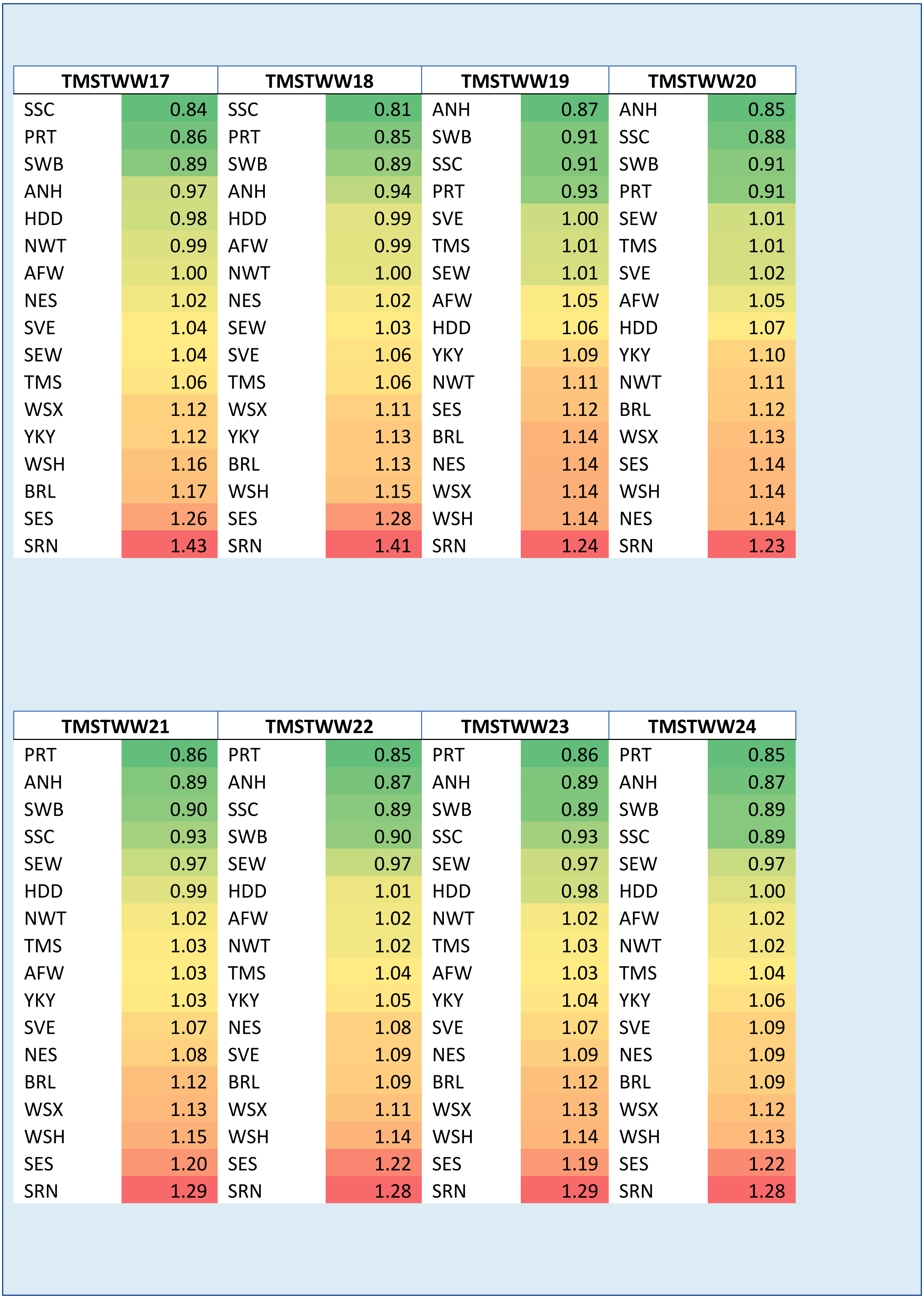
Efficiency scores distribution

TMSTWW1		TMSTWW2		TMSTWW3		TMSTWW4	
SSC	0.81	SSC	0.78	PRT	0.83	SSC	0.80
PRT	0.94	PRT	0.92	SSC	0.84	PRT	0.82
SVE	0.95	ANH	0.93	NWT	0.97	ANH	0.95
ANH	0.95	SVE	0.97	AFW	0.98	NWT	0.98
SWB	1.00	SWB	1.00	NES	0.99	AFW	0.98
TMS	1.01	AFW	1.01	ANH	0.99	SWB	0.99
AFW	1.02	TMS	1.03	SWB	0.99	SEW	0.99
SEW	1.05	SEW	1.04	SEW	1.00	NES	0.99
NES	1.07	NES	1.07	SVE	1.04	TMS	1.05
NWT	1.10	WSX	1.09	TMS	1.04	SVE	1.06
WSX	1.10	NWT	1.10	YKY	1.05	YKY	1.06
YKY	1.12	YKY	1.13	WSX	1.09	WSX	1.08
SES	1.18	BRL	1.17	HDD	1.10	HDD	1.12
HDD	1.19	SES	1.20	BRL	1.19	BRL	1.15
BRL	1.20	HDD	1.21	WSH	1.24	WSH	1.22
WSH	1.23	WSH	1.22	SES	1.30	SES	1.33
SRN	1.33	SRN	1.32	SRN	1.41	SRN	1.39

TMSTWW5		TMSTWW6		TMSTWW7		TMSTWW8	
SSC	0.83	SSC	0.80	ANH	0.88	SSC	0.86
PRT	0.84	PRT	0.83	SSC	0.89	ANH	0.86
NWT	0.97	ANH	0.95	PRT	0.92	PRT	0.90
SWB	0.98	SWB	0.97	SEW	0.97	SEW	0.97
AFW	0.99	AFW	0.98	TMS	0.99	SWB	1.00
ANH	0.99	NWT	0.98	SVE	0.99	TMS	1.00
NES	1.00	SEW	0.99	SWB	1.00	SVE	1.01
SEW	1.00	NES	1.00	AFW	1.03	AFW	1.03
SVE	1.04	TMS	1.05	YKY	1.03	YKY	1.05
TMS	1.04	SVE	1.06	NWT	1.10	WSX	1.09
YKY	1.06	YKY	1.07	WSX	1.11	NWT	1.10
HDD	1.07	WSX	1.08	NES	1.12	BRL	1.12
WSX	1.09	HDD	1.09	BRL	1.13	NES	1.12
BRL	1.18	BRL	1.14	SES	1.15	SES	1.17
WSH	1.23	WSH	1.21	HDD	1.17	HDD	1.19
SES	1.30	SES	1.32	WSH	1.21	WSH	1.20
SRN	1.42	SRN	1.39	SRN	1.22	SRN	1.21

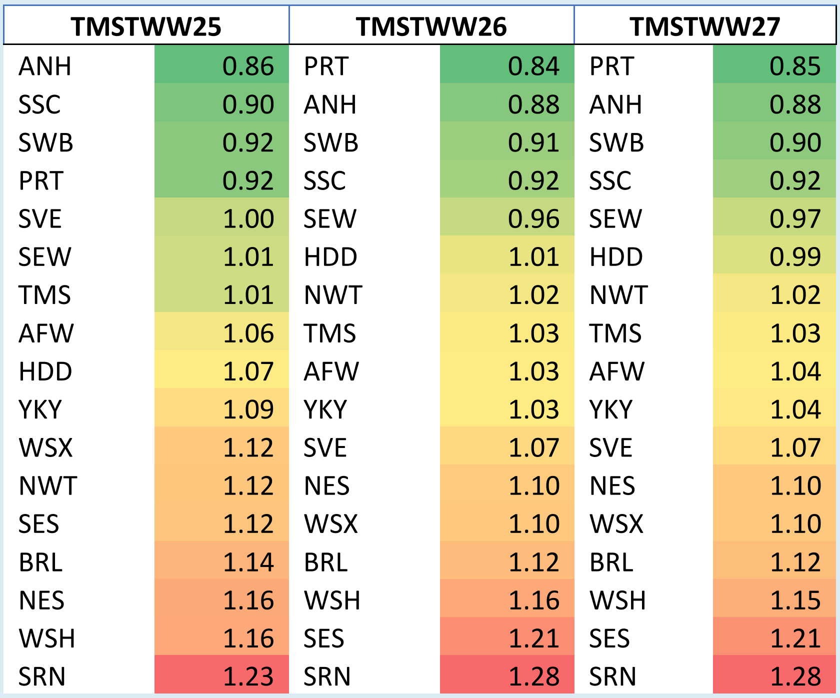
TMSTWW9		TMSTWW10		TMSTWW11		TMSTWW12	
PRT	0.83	PRT	0.82	PRT	0.84	PRT	0.83
ANH	0.90	ANH	0.88	ANH	0.90	SSC	0.88
SSC	0.93	SSC	0.89	SSC	0.92	ANH	0.88
SEW	0.93	SEW	0.93	SEW	0.94	SEW	0.93
YKY	0.98	SWB	0.99	SWB	0.98	SWB	0.98
SWB	0.99	YKY	1.00	YKY	0.99	YKY	1.00
NWT	1.00	NWT	1.00	NWT	1.00	AFW	1.00
AFW	1.01	AFW	1.01	AFW	1.01	NWT	1.01
TMS	1.02	TMS	1.03	TMS	1.02	TMS	1.03
NES	1.06	NES	1.06	NES	1.06	NES	1.06
SVE	1.07	WSX	1.08	SVE	1.07	WSX	1.08
HDD	1.09	SVE	1.09	HDD	1.08	SVE	1.09
WSX	1.10	BRL	1.10	WSX	1.10	BRL	1.10
BRL	1.13	HDD	1.12	BRL	1.12	HDD	1.10
WSH	1.21	WSH	1.20	WSH	1.20	WSH	1.20
SES	1.23	SRN	1.26	SES	1.23	SES	1.26
SRN	1.27	SES	1.26	SRN	1.28	SRN	1.26

TMSTWW13		TMSTWW14		TMSTWW15		TMSTWW16	
SSC	0.82	SSC	0.80	SSC	0.84	SSC	0.81
SWB	0.91	SWB	0.91	PRT	0.86	PRT	0.85
ANH	0.94	ANH	0.92	SWB	0.90	SWB	0.90
PRT	0.95	PRT	0.93	ANH	0.98	ANH	0.94
SVE	0.95	SVE	0.98	NWT	0.99	AFW	0.99
TMS	1.03	AFW	1.03	HDD	1.00	NWT	0.99
AFW	1.04	TMS	1.04	AFW	1.00	HDD	1.01
HDD	1.08	NES	1.08	NES	1.01	NES	1.02
NES	1.09	SEW	1.09	SEW	1.03	SEW	1.03
SEW	1.09	HDD	1.09	SVE	1.04	SVE	1.06
NWT	1.11	NWT	1.11	TMS	1.05	TMS	1.06
WSX	1.14	WSX	1.13	YKY	1.11	WSX	1.11
SES	1.14	WSH	1.15	WSX	1.11	YKY	1.12
WSH	1.17	SES	1.16	BRL	1.17	BRL	1.14
YKY	1.18	BRL	1.17	WSH	1.18	WSH	1.16
BRL	1.20	YKY	1.19	SES	1.26	SES	1.29
SRN	1.36	SRN	1.34	SRN	1.43	SRN	1.41



TMSTWW17		TMSTWW18		TMSTWW19		TMSTWW20	
SSC	0.84	SSC	0.81	ANH	0.87	ANH	0.85
PRT	0.86	PRT	0.85	SWB	0.91	SSC	0.88
SWB	0.89	SWB	0.89	SSC	0.91	SWB	0.91
ANH	0.97	ANH	0.94	PRT	0.93	PRT	0.91
HDD	0.98	HDD	0.99	SVE	1.00	SEW	1.01
NWT	0.99	AFW	0.99	TMS	1.01	TMS	1.01
AFW	1.00	NWT	1.00	SEW	1.01	SVE	1.02
NES	1.02	NES	1.02	AFW	1.05	AFW	1.05
SVE	1.04	SEW	1.03	HDD	1.06	HDD	1.07
SEW	1.04	SVE	1.06	YKY	1.09	YKY	1.10
TMS	1.06	TMS	1.06	NWT	1.11	NWT	1.11
WSX	1.12	WSX	1.11	SES	1.12	BRL	1.12
YKY	1.12	YKY	1.13	BRL	1.14	WSX	1.13
WSH	1.16	BRL	1.13	NES	1.14	SES	1.14
BRL	1.17	WSH	1.15	WSX	1.14	WSH	1.14
SES	1.26	SES	1.28	WSH	1.14	NES	1.14
SRN	1.43	SRN	1.41	SRN	1.24	SRN	1.23

TMSTWW21		TMSTWW22		TMSTWW23		TMSTWW24	
PRT	0.86	PRT	0.85	PRT	0.86	PRT	0.85
ANH	0.89	ANH	0.87	ANH	0.89	ANH	0.87
SWB	0.90	SSC	0.89	SWB	0.89	SWB	0.89
SSC	0.93	SWB	0.90	SSC	0.93	SSC	0.89
SEW	0.97	SEW	0.97	SEW	0.97	SEW	0.97
HDD	0.99	HDD	1.01	HDD	0.98	HDD	1.00
NWT	1.02	AFW	1.02	NWT	1.02	AFW	1.02
TMS	1.03	NWT	1.02	TMS	1.03	NWT	1.02
AFW	1.03	TMS	1.04	AFW	1.03	TMS	1.04
YKY	1.03	YKY	1.05	YKY	1.04	YKY	1.06
SVE	1.07	NES	1.08	SVE	1.07	SVE	1.09
NES	1.08	SVE	1.09	NES	1.09	NES	1.09
BRL	1.12	BRL	1.09	BRL	1.12	BRL	1.09
WSX	1.13	WSX	1.11	WSX	1.13	WSX	1.12
WSH	1.15	WSH	1.14	WSH	1.14	WSH	1.13
SES	1.20	SES	1.22	SES	1.19	SES	1.22
SRN	1.29	SRN	1.28	SRN	1.29	SRN	1.28



TMSTWW25		TMSTWW26		TMSTWW27	
ANH	0.86	PRT	0.84	PRT	0.85
SSC	0.90	ANH	0.88	ANH	0.88
SWB	0.92	SWB	0.91	SWB	0.90
PRT	0.92	SSC	0.92	SSC	0.92
SVE	1.00	SEW	0.96	SEW	0.97
SEW	1.01	HDD	1.01	HDD	0.99
TMS	1.01	NWT	1.02	NWT	1.02
AFW	1.06	TMS	1.03	TMS	1.03
HDD	1.07	AFW	1.03	AFW	1.04
YKY	1.09	YKY	1.03	YKY	1.04
WSX	1.12	SVE	1.07	SVE	1.07
NWT	1.12	NES	1.10	NES	1.10
SES	1.12	WSX	1.10	WSX	1.10
BRL	1.14	BRL	1.12	BRL	1.12
NES	1.16	WSH	1.16	WSH	1.15
WSH	1.16	SES	1.21	SES	1.21
SRN	1.23	SRN	1.28	SRN	1.28

Comments

- Please indicate the units of the explanatory variable, and whether it was expressed in logs.
- Use asterisks to denote significance level: *** (1%), ** (5%) and * (10%)
- P values should be based on cluster robust standard errors
- In the case of random effects please report Stata's output "R2 overall"
- Please use the following naming convention to assign a name to each model: company acronym, level of aggregation, model number (eg for Anglian Water's wholesale water model number 1: ANHWW1). Please refer to the table below for company acronyms and level of aggregation acronyms.

Company acronyms	Level of aggregation acronyms
Anglian Water: ANH Hafren Dyfrdwy: HDD Northumbrian Water: NES Southern Water: SRN Severn Trent England: SVE South West Water: SWB Thames Water: TMS United Utilities: UUW Dŵr Cymru: WSH Wessex Water: WSX Yorkshire Water: YKY Affinity Water: AFW Bristol Water: BRL Portsmouth Water: PRT SES Water: SES South East Water: SEW South Staffs Water: SSC	**Wholesale water** Treated water distribution: TWD Water resources plus: WRP Water network plus: WWNP Wholesale water: WW **Wholesale wastewater** Sewage collection: SWC Sewage treatment: STW Bioresources: BR Wastewater network plus: WWWNP Bioresources plus: BRP **Residential retail** Bad debt related costs: RDC Other costs: ROC Total costs: RTC

Template for submission of econometric models for consultation (Sewage Collection (SWC))

Econometric model formula:

1. TMSSWC1: ln(WWWNP botex plus Network Reinforcement and Reduced Sewer flooding Growth $_{it}$) = $\alpha + \beta_1$ ln(Sewer Length $_{it}$) + β_2 ln(Pumping Capacity per Length $_{it}$) + β_3 ln(weighted average density LAD $_{it}$) + β_4 (ln(weighted average density LAD $_{it}$))^2 + β_5 Ln(Urban Rainfall LAD $_{it}$) + ε_{it}

2. TMSSWC2: ln(WWWNP botex plus Network Reinforcement and Reduced Sewer flooding Growth $_{it}$) = $\alpha + \beta_1$ ln(Sewer Length $_{it}$) + β_2 ln(Pumping Capacity per Length $_{it}$) + β_3 ln(weighted average density MSOA population $_{it}$) + β_4 (ln(weighted average density MSOA population $_{it}$))^2 + β_5 Ln(Urban Rainfall LAD $_{it}$) + ε_{it}

3. TMSSWC3: ln(WWWNP botex plus Network Reinforcement and Reduced Sewer flooding Growth $_{it}$) = $\alpha + \beta_1$ ln(Sewer Length $_{it}$) + β_2 ln(Pumping Capacity per Length $_{it}$) + β_3 ln(weighted average density LAD $_{it}$) + β_4 (ln(weighted average density LAD $_{it}$))^2 + β_5 Ln(Urban Rainfall MOSA $_{it}$) + ε_{it}

4. TMSSWC4: ln(WWWNP botex plus Network Reinforcement and Reduced Sewer flooding Growth $_{it}$) = $\alpha + \beta_1$ ln(Sewer Length $_{it}$) + β_2 ln(Pumping Capacity per Length $_{it}$) + β_3 ln(weighted average density LAD $_{it}$) + β_4 (ln(weighted average density LAD $_{it}$))^2 + β_5 Ln(Annual Rainfall $_{it}$) + ε_{it}

5. TMSSWC5: ln(WWWNP botex plus Network Reinforcement and Reduced Sewer flooding Growth $_{it}$) = $\alpha + \beta_1$ ln(Sewer Length $_{it}$) + β_2 ln(Pumping Capacity per Length $_{it}$) + β_3 ln(weighted average density MSOA population $_{it}$) + β_4 (ln(weighted average density MSOA population $_{it}$))^2 + β_5 Ln(Annual Rainfall $_{it}$) + ε_{it}

6. TMSSWC6: ln(WWWNP botex plus Network Reinforcement and Reduced Sewer flooding Growth $_{it}$) = $\alpha + \beta_1$ ln(Composite Scale Variable (CSV) $_{it}$) + β_2 ln(Pumping Capacity per Length $_{it}$) + β_3 ln(weighted average density LAD $_{it}$) + β_4 (ln(weighted average density LAD $_{it}$))^2 + ε_{it}

7. TMSSWC7: ln(WWWNP botex plus Network Reinforcement and Reduced Sewer flooding Growth $_{it}$) = $\alpha + \beta_1$ ln(Composite Scale Variable (CSV) $_{it}$) + β_2 ln(Pumping Capacity per Length $_{it}$) + β_3 ln(weighted average density LAD $_{it}$) + β_4 (ln(weighted average density LAD $_{it}$))^2 + β_5 Ln(Urban Rainfall LAD $_{it}$) + ε_{it}

8. TMSSWC8: ln(WWWNP botex plus Network Reinforcement and Reduced Sewer flooding Growth $_{it}$) = $\alpha + \beta_1$ ln(Composite Scale Variable (CSV) $_{it}$) + β_2 ln(Pumping Capacity per Length $_{it}$) + β_3 ln(weighted average density MSOA population $_{it}$) + β_4 (ln(weighted average density MSOA population $_{it}$))^2 + β_5 Ln(Urban Rainfall LAD $_{it}$) + ε_{it}

9. TMSSWC9: ln(WWWNP botex plus Network Reinforcement and Reduced Sewer flooding Growth $_{it}$) = $\alpha + \beta_1$ ln(Composite Scale Variable (CSV) $_{it}$) + β_2 ln(Pumping Capacity per Length $_{it}$) + β_3 ln(weighted average density LAD $_{it}$) + β_4 (ln(weighted average density LAD $_{it}$))^2 + β_5 Ln(Annual Rainfall $_{it}$) + ε_{it}

10. TMSSWC10: ln(WWWNP botex plus Network Reinforcement and Reduced Sewer flooding Growth $_{it}$) = $\alpha + \beta_1$ ln(Composite Scale Variable (CSV) $_{it}$) + β_2 ln(Pumping Capacity per Length $_{it}$) + β_3 ln(weighted average density MSOA population $_{it}$) + β_4 (ln(weighted average density MSOA population $_{it}$))^2 + β_5 Ln(Annual Rainfall $_{it}$) + ε_{it}

Description of the dependent variable

All the econometric models presented in this template use the same definition of **Botex Plus for Wholesale Wastewater Network Plus (Collection and Treatment)** as defined by Ofwat in the Stata code below (e.g., see below **botexplusnpww**):

Sewage Collection Botex

```
g botexswc = BM402SC + BM836SC + BM431SC + BM140SC + BM839ISC + BM839NISC + BM839OSC
  + BC30945SC + CS00036SC + S3024SC + BN4012_SWC - W3032NPSC - W3036NPSC - APP28RR_WW0002 - APP28RR_WW0003 - B0201DSWWADJ
```

Sewage Treatment Botex

```
g botexswt = BM502ST + BM836ST + BM531ST
  + BM140ST + BM839IST + BM839NIST + BM839OST
  + BC30945ST + CS00036ST + S3024ST + BN4012_SWT - W3032NPST - W3036NPST
- BN5000 + B0312CRO_SWT + B0318NRO_SWT + B0321PRO_SWT + B0324RSO_SWT + B0327UVO_SWT
```

Sewage Collection Botex Plus

```
g botexplusswc = botexswc + S3023SC + B0337RFO_TOT + B0200DSISWCWWC + B0200DSISWCWWO
```

Sewage Treatment Botex Plus

```
g botexplusswt = botexswt + S3023ST
```

Wholesale Wastewater Network Plus Botex

```
g botexnpww = botexswc + botexswt
```

Botex Plus Wholesale Wastewater Network Plus (Botex Plus Network Reinforcement and Reduced Sewer flooding Growth lines for SWC and SWT):

```
g botexplusnpww = botexplusswc + botexplusswt
```

Description of the explanatory variables

- **Ln(Load) = Natural Log or Ln(Load) = Ln(STWDP125_21);** kg BOD5/day
- **Ln(Properties) = Natural Log or Ln(Properties) = Ln((BN1178) * 1000);** Number of Properties
- **Length of Sewage: sewerlength** = BN13535_21 + BN13528 ; Km
- **Composite Scale Variable (for Sewage Collection) (CSV) = (sewerlength)^0.5 * (Properties)^0.5; We allocate the same weight to sewage length and properties, although these weights are flexible to other figures.**
- **Ln(CSV)= Natural Log or Ln(CSV)**
- **Ln(Pumping Capacity per Length)** = Ln(S4029 / sewerlength); (kW/Km)
- **Ln(Urban Rainfall LAD)** = Ln(urban_rainfall_lad)
- **Ln(Urban Rainfall MOSA)** = Ln(urban_ urban_rainfall_mosa)
- **Ln(Annual Rainfall) =** Ln(annual_rainfall)

Brief comment on the models

- The aim of the models presented in this section for SWC is to provide evidence on how the current models in PR19 could be improved. We find that the use of average property density as a cost driver (Properties/Mains) does not really reflect the different levels of density faced by each company across the industry. We believe that Weighted Average Population Density LAD and MSOA are good complements to show the effect of density in the industry. Moreover, we have found that the rainfall effect on base cost is robust and significant. In particular, the effect of urban rainfall seems to provide more confident results when compared to the other two alternatives of measuring rainfall. For example, the standard errors are lower and the R^2 higher when using urban rainfall lad as a cost driver. Lastly, we introduce the potential use of a Composite Scale Variable (CSV) for SWC models. The results indicate improvements in the models, but more discussion on the weights of the drivers used in the CSV calculation is needed.
- All the models are run using the period of 11 years, 2011-12 to 2021-22.
- All our proposed models in SWC improve the R^2 when compared with the PR19 SWC2 model as we are proposing models with the Weighted Average Population Density (LAD or MOSA) for comparison purposes. Moreover, our proposed models TMSSWC1-

3, 7-8 are quite robust to all the statistical checks and changes to the removing of years or (in)efficient companies.

- Our models use Weighted Average Population Density (LAD and MSOA) as we believe this to be the most appropriate representation of density in the SWC business unit. We consider that the current property average density (Properties/Mains) used in one of the two PR19 models should be substituted with the Weighted Average Population Density (LAD and MOSA). The reasons for this are:
    - i) We consider that *population density* is more beneficial to understanding operating costs than property density, for the simple reason that it is the wastewater produced (load) by people that generates the cost to operate. Variances in population density versus property density per sq km take into account not only the property type i.e., 1 bedroom starter homes versus maisonettes, flats or large multi bedroom houses where the occupancy number will be greater but also a bit more about the demographics i.e., areas may vary in terms of occupancy based on possibly house or location value where it is more common for single occupancy of homes versus multiple occupancy due to individuals personal circumstances. For example, in a wastewater hydraulic modelling the approach is always based on occupancy (population density) as it is not possible to derive wastewater usage profiles from just a house (property) count.
    - ii) The property average density (Properties/Mains) could be replaced with the Weighted Average Density (MSOA) proposed in the dataset, for instance see models TMSSWC 2,5,8 and 10. Using the LAD and MSOA density drivers provides consistency in the structure of the models used and with the *population density* definition that is also used in water. We believe that this could be a robust way to show the effect of density in SWC.
    - iii) The model that includes average property density (Properties/Mains) in PR19 does not show any significant effect when its corresponding square term is used, in other words the coefficient of average density (Properties/Mains) and its square term are insignificant. This suggests some inconsistency with the other model that uses Weighted Average Density (LAD and MOSA) and its square term as the one proposed by the CMA. The square term is quite important in the industry to distinguish the different levels of density that each company faces
    - iv) We believe that for consistency with the water models, the SCW models should be aligned with Population Density using the Weighted Average Density (LAD/MOSA). In addition, this measure provides a weighted average rather than a simple one, reflecting a stronger link between costs and density.

- In all our proposed models all the cost drivers are statistically significant even when new drivers are proposed in the models as is the case of the rainfall variables or the Composite scale variable (CSV).
- Most of our proposed models are robust based on the guidance proposed by Ofwat, as can be appreciated in the table results.
- We have found a significant effect of the rainfall variables proposed in the dataset. This follows our suggestion as a cost driver in our December 2021 "*Assessing Base Cost at PR24*" consultation response. In that response we provide some suggestions on why this is a good driver for sewage collection. So far, the results in our proposed models in this template are align with the insights proposed in our response in 2021 Base consultation. When rainfall is included the performance of the R^2 and the robustness of the models is improved alongside with population density and its square term to capture the different levels of density faced by waste companies. The rainfall driver could yield a proxy for the amount of rainfall that wastewater companies need to deal with every year on historical base.
- The results in the models also suggest that urban rainfall seems to perform or reflect a stronger link with base costs than when annual rainfall is used. Nevertheless, both measures are demonstrated to be robust drivers of cost, regardless of which is used.
- Finally, we also introduce the effect of a Composite Scale Variable comprising two measures of scale: Sewage length of mains and Number of Properties. We consider that these two dimensions of scale are a good representation for a SWC model as they are closer to the network characteristics of this part of the value chain. At this stage, we are agnostic as to the relative weight of the drivers and have therefore assigned an equal 50% weight to each. Further work could be done to refine the weights to best represent the specifics of sewage collection systems. Including the CSV measure improves the R^2 in all cases but note that models using only sewage length are also reasonably robust and less sensitive to removing either years from the sample or the most/less efficient company of the industry (see models TWSSWC1-5). Moreover, using length of sewage could mitigate uncertainty in the weights within the CSV measure.

	TMSSWC1	TMSSWC2	TMSSWC3	TMSSWC4
Dependent Variable	Botex + Network Reinforcement & Reduced Sewer flooding Growth	Botex + Network Reinforcement & Reduced Sewer flooding Growth	Botex + Network Reinforcement & Reduced Sewer flooding Growth	Botex + Network Reinforcement & Reduced Sewer flooding Growth
Ln(Sewer Length)	0.704*** (0.000)	0.705*** (0.000)	0.710*** (0.000)	0.899*** (0.000)
Ln(Pumping Capacity per Length)	0.563*** (0.000)	0.515*** (0.000)	0.574*** (0.000)	0.621*** (0.000)
Ln(WAD_LAD)	-2.147*** (0.000)		-2.200*** (0.000)	-2.297*** (0.003)
(Ln(WAD_LAD))^2	0.161*** (0.000)		0.164*** (0.000)	0.169*** (0.001)
Ln(Urban Rainfall LAD)	0.167*** (0.000)	0.168*** (0.000)		
Ln(WAD_MSOA_population)		-4.854*** (0.001)		
(Ln(WAD_MSOA_population))^2		0.324*** (0.000)		
Ln(Urban Rainfall MOSA)			0.161*** (0.000)	
Ln(Annual Rainfall)				0.145*** (0.000)
Ln(CSV)				
Constant	2.741 (0.129)	13.761** (0.020)	2.9 (0.138)	1.484 (0.611)
Estimation Method (OLS or RE)	RE	RE	RE	RE
N (sample size)	110	110	110	110
Model robustness tests				
R2 adjusted	0.923	0.923	0.920	0.909
RESET test	0.691	0.781	0.663	0.690
VIF (max) (OLS)	400	984	400	404
Pooling / Chow test (OLS)	0.914	0.900	0.951	0.997

Normality of model residuals (OLS)	0.001	0.002	0.002	0.004
Heteroskedasticity of model residuals (OLS)	0.005	0.004	0.003	0.001
Test of pooled OLS versus Random Effects (LM test)	0.000	0.000	0.000	0.000
Efficiency Score Distribution	Min: 0.90	Min: 0.89	Min: 0.89	Min: 0.89
	Max: 1.18	Max: 1.13	Max: 1.18	Max: 1.21
Sensitivity of estimated coefficients to removal of most and least efficient company	G	G	G	G
Sensitivity of estimated coefficients to removal of first and last year of the sample	G	G	G	A

	TMSSWC5	TMSSWC6	TMSSWC7	TMSSWC8
Dependent Variable	Botex + Network Reinforcement & Reduced Sewer flooding Growth	Botex + Network Reinforcement & Reduced Sewer flooding Growth	Botex + Network Reinforcement & Reduced Sewer flooding Growth	Botex + Network Reinforcement & Reduced Sewer flooding Growth
Ln(Sewer Length)	0.892*** (0.000)			
Ln(Pumping Capacity per Length)	0.568*** (0.000)	0.501*** (0.000)	0.483*** (0.000)	0.452*** (0.000)
Ln(WAD_LAD)		-2.067*** (0.009)	-1.935*** (0.000)	
(Ln(WAD_LAD))^2		0.148*** (0.005)	0.142*** (0.000)	
Ln(Urban Rainfall LAD)			0.152*** (0.000)	0.154*** (0.000)
Ln(WAD_MSOA_population)	-4.559** (0.033)			-4.822*** (0.000)
(Ln(WAD_MSOA_population))^2	0.305** (0.019)			0.315*** (0.000)
Ln(Urban Rainfall MOSA)				
Ln(Annual Rainfall)	0.148*** (0.000)			
Ln(CSV)		0.839*** (0.000)	0.693*** (0.000)	0.699*** (0.000)
Constant	10.79 (0.205)	1.094 (0.719)	1.177 (0.384)	12.964*** (0.002)
Estimation Method (OLS or RE)	RE	RE	RE	RE
N (sample size)	110	110	110	110
Model robustness tests				
R2 adjusted	0.911	0.916	0.931	0.931
RESET test	0.886	0.351	0.540	0.638
VIF (max) (OLS)	974	402	403	984
Pooling / Chow test (OLS)	0.994	0.908	0.879	0.842
Normality of model residuals (OLS)	0.006	0.065	0.001	0.001

Heteroskedasticity of model residuals (OLS)	0.001	0.155	0.020	0.014
Test of pooled OLS versus Random Effects (LM test)	0.000	0.000	0.002	0.001
Efficiency Score Distribution	Min: 0.87	Min: 0.90	Min: 0.92	Min: 0.92
	Max: 1.17	Max: 1.14	Max: 1.13	Max: 1.10
Sensitivity of estimated coefficients to removal of most and least efficient company	A	A	G	G
Sensitivity of estimated coefficients to removal of first and last year of the sample	A	G	G	G

	TMSSWC9	TMSSWC10
Dependent Variable	Botex + Network Reinforcement & Reduced Sewer flooding Growth	Botex + Network Reinforcement & Reduced Sewer flooding Growth
Ln(Sewer Length)		
Ln(Pumping Capacity per Length)	0.510*** 0.000	0.483*** 0.000
Ln(WAD_LAD)	-1.793*** 0.000	
(Ln(WAD_LAD))^2	0.129*** 0.000	
Ln(Urban Rainfall LAD)		
Ln(WAD_MSOA_population)		-4.087*** 0.001
(Ln(WAD_MSOA_population))^2		0.266*** 0.000

Ln(Urban Rainfall MOSA)		
Ln(Annual Rainfall)	0.153*** 0.000	0.155*** 0.000
Ln(CSV)	0.878*** 0.000	0.881*** 0.000
Constant	-1.465 0.31	7.979* 0.08
Estimation Method (OLS or RE)	RE	RE
N (sample size)	110	110
Model robustness tests		
R2 adjusted	0.929	0.929
RESET test	0.767	0.844
VIF (max) (OLS)	407	973
Pooling / Chow test (OLS)	0.936	0.915
Normality of model residuals (OLS)	0.000	0.001
Heteroskedasticity of model residuals (OLS)	0.012	0.009
Test of pooled OLS versus Random Effects (LM test)	0.000	0.000
Efficiency Score Distribution	Min: 0.92	Min: 0.91
	Max: 1.14	Max: 1.12
Sensitivity of estimated coefficients to removal of most and least efficient company	G	G
Sensitivity of estimated coefficients to removal of first and last year of the sample	A	A

Efficiency scores distribution

TMSSWC1		TMSSWC2		TMSSWC3		TMSSWC4	
WSX	0.90	WSX	0.89	WSX	0.89	WSX	0.89
ANH	0.96	ANH	0.96	ANH	0.96	ANH	0.92
WSH	0.97	TMS	0.98	WSH	0.97	SWB	0.95
SRN	1.00	WSH	0.99	TMS	1.00	TMS	0.97
TMS	1.00	SRN	1.00	SRN	1.00	WSH	1.00
SWB	1.01	NWT	1.01	SWB	1.01	SRN	1.01
NWT	1.01	SWB	1.02	NWT	1.02	NWT	1.03
SVH	1.04	NES	1.05	SVH	1.04	SVH	1.05
NES	1.05	SVH	1.07	NES	1.06	NES	1.08
YKY	1.18	YKY	1.13	YKY	1.18	YKY	1.21

TMSSWC5		TMSSWC6		TMSSWC7		TMSSWC8	
WSX	0.87	WSX	0.90	WSX	0.92	WSX	0.92
ANH	0.94	ANH	0.90	WSH	0.95	WSH	0.96
TMS	0.96	TMS	0.99	ANH	0.98	ANH	0.97
SWB	0.96	SRN	0.99	SRN	0.99	SRN	1.00
SRN	1.01	NES	1.01	TMS	1.01	TMS	1.00
NWT	1.02	WSH	1.01	NES	1.02	NES	1.02
WSH	1.04	SVH	1.02	SVH	1.02	NWT	1.02
NES	1.06	SWB	1.02	NWT	1.02	SVH	1.04
SVH	1.07	NWT	1.11	SWB	1.05	SWB	1.05
YKY	1.16	YKY	1.14	YKY	1.13	YKY	1.10

TMSSWC9		TMSSWC10	
WSX	0.92	WSX	0.91
ANH	0.95	ANH	0.95
WSH	0.97	WSH	0.99
SRN	1.00	TMS	0.99
SWB	1.00	SWB	1.00
TMS	1.01	SRN	1.00
SVH	1.02	NWT	1.02
NWT	1.02	NES	1.03
NES	1.03	SVH	1.03
YKY	1.14	YKY	1.12

Comments

- Please indicate the units of the explanatory variable, and whether it was expressed in logs.
- Use asterisks to denote significance level: *** (1%), ** (5%) and * (10%)
- P values should be based on cluster robust standard errors
- In the case of random effects please report Stata's output "R2 overall"
- Please use the following naming convention to assign a name to each model: company acronym, level of aggregation, model number (eg for Anglian Water's wholesale water model number 1: ANHWW1). Please refer to the table below for company acronyms and level of aggregation acronyms.

Company acronyms	Level of aggregation acronyms
Anglian Water: ANH Hafren Dyfrdwy: HDD Northumbrian Water: NES Southern Water: SRN Severn Trent England: SVE South West Water: SWB Thames Water: TMS United Utilities: UUW Dŵr Cymru: WSH Wessex Water: WSX Yorkshire Water: YKY Affinity Water: AFW Bristol Water: BRL Portsmouth Water: PRT SES Water: SES South East Water: SEW South Staffs Water: SSC	**Wholesale water** Treated water distribution: TWD Water resources plus: WRP Water network plus: WWNP Wholesale water: WW **Wholesale wastewater** Sewage collection: SWC Sewage treatment: STW Bioresources: BR Wastewater network plus: WWWNP Bioresources plus: BRP **Residential retail** Bad debt related costs: RDC Other costs: ROC Total costs: RTC

Template for submission of econometric models for consultation (Sewage Treatment)

Econometric model formula:

1. TMSSWT1: ln(WWWNP botex plus Network Reinforcement and Reduced Sewer flooding Growth $_{it}$) = α + β_1 ln(Load$_{it}$) + β_2 (PCT NH3 below 3mg$_{it}$) + β_3 (PCT Bands 1-3 $_{it}$) + β_4 ln(Pumping Capacity per Length $_{it}$) + ε_{it}

2. TMSSWT2: ln(WWWNP botex plus Network Reinforcement and Reduced Sewer flooding Growth $_{it}$) = α + β_1 ln(Load$_{it}$) + β_2 (PCT NH3 below 3mg$_{it}$) + β_3 (PCT Bands 6 $_{it}$) + β_4 ln(Pumping Capacity per Length $_{it}$) + ε_{it}

Description of the dependent variable

All the econometric models presented in this template use the same definition of **Botex Plus for Wholesale Wastewater Network Plus (Collection and Treatment)** as defined by Ofwat in the Stata code below (e.g., see below **botexplusnpww**):

Sewage Collection Botex

g **botexswc** = BM402SC + BM836SC + BM431SC + BM140SC + BM839ISC + BM839NISC + BM839OSC + BC30945SC + CS00036SC + S3024SC + BN4012_SWC – W3032NPSC – W3036NPSC – APP28RR_WW0002 – APP28RR_WW0003 – B0201DSWWADJ

Sewage Treatment Botex

g **botexswt** = BM502ST + BM836ST + BM531ST + BM140ST + BM839IST + BM839NIST + BM839OST + BC30945ST + CS00036ST + S3024ST + BN4012_SWT – W3032NPST – W3036NPST – BN5000 + B0312CRO_SWT + B0318NRO_SWT + B0321PRO_SWT + B0324RSO_SWT + B0327UVO_SWT

Sewage Collection Botex Plus

g **botexplusswc** = botexswc + S3023SC + B0337RFO_TOT + B0200DSISWCWWC + B0200DSISWCWWO

Sewage Treatment Botex Plus

g **botexplusswt** = botexswt + S3023ST

Wholesale Wastewater Network Plus Botex

g **botexnpww** = botexswc + botexswt

Botex Plus Wholesale Wastewater Network Plus (Botex Plus Network Reinforcement and Reduced Sewer flooding Growth lines for SWC and SWT):

g **botexplusnpww** = botexplusswc + botexplusswt

Description of the explanatory variables

- **Ln(Load) = Natural Log or Ln(Load) = Ln(STWDP125_21);** kg BOD5/day
- **Ln(Properties) = Natural Log or Ln(Properties) = Ln((BN1178) * 1000);** Number of Properties
- **Length of Sewage: sewerlength** = BN13535_21 + BN13528; Km
- **Ln(Pumping Capacity per Length)** = Ln(S4029 / sewerlength); (kW/Km)
- **PCT NH3 below 3mg** = ((STWDA121) / (load)) * 100; %
- **PCT bands 1-3** = ((STWDP005_21 + STWDP019_21 + STWDP033_21) / (load)) * 100; %
- **PCT bands 6** = ((STWDP105_21) / (load)) * 100; %

Brief comment on the models

- The SWT models proposed in this section provide potential alternatives to the current SWT models. Our models TMSSWT1-2, provide an improvement on the R^2 when compared to the current PR19 model SWT2. This is because of the inclusion of Pumping Capacity per Main. Sewage Pumping stations capacity can provide a good level of insight into the operation of a Sewage Treatment Works (SWTs). In general, STWs will either receive incoming flow via gravity, pumped flows, or a combination of the two. For those sites where the dominant flow is pumped, correlation between the operation of the pumping station and the STWs can be hugely insightful, typically this will involve looking at the terminal pumping stations only i.e., those pumping stations that outfall directly to the STWs. The pumping station capacity can be helpful because it provides insight on the total flow passed to the STWs and not just the treated flow. Total flow will pass through screens etc. capturing costs that might not otherwise be allowed for.
- Model TMSSWT1 also improves the fitness of the model but struggles with the RESET-test.
- All the models are run using the period of 11 years, 2011-12 to 2021-22.
- The models TMSSWT1-2 have an R^2 that is higher than the current PR19 models, with an R^2 of 0.89, although model TMSSWT3 struggles with the RESET-test.
- Overall, these models reduce the spread of the efficiency scores when compared with the PR19 results, which provides more confidence at the industry level on the base costs that are being explained.

	TMSSWT1	TMSSWT2
Dependent Variable	Botex + Network Reinforcement & Reduced Sewer flooding Growth	Botex + Network Reinforcement & Reduced Sewer flooding Growth
Ln(Load)	0.748*** 0.000	0.736*** 0.000
PCT NH3 below 3mg	0.005*** 0.000	0.006*** 0.000
Ln(WAD_LAD)		
Ln(Pumping Capacity per Length)	0.333* 0.08	0.322 0.113
PCT Bands 1-3	0.034* 0.064	
PCT Bands 6		-0.010** 0.043
Constant	-5.082*** 0.000	-4.051*** 0.000
Estimation Method (OLS or RE)	RE	RE
N (sample size)	110	110
Model robustness tests		
R2 adjusted	0.891	0.891
RESET test	0.025	0.178
VIF (max) (OLS)	5.396	4.453

Pooling / Chow test (OLS)	1.000	1.000
Normality of model residuals (OLS)	0.022	0.012
Heteroskedasticity of model residuals (OLS)	0.014	0.083
Test of pooled OLS versus Random Effects (LM test)	0.000	0.000
Efficiency Score Distribution	Min: 0.83	Min: 0.86
	Max: 1.17	Max: 1.20
Sensitivity of estimated coefficients to removal of most and least efficient company	A	A
Sensitivity of estimated coefficients to removal of first and last year of the sample	A	A

Efficiency scores distribution

TWSSWT1		TWSSWT2	
TMS	0.83	TMS	0.86
SWB	0.94	SVH	0.98
SVH	0.98	WSX	0.98
NES	1.00	SWB	0.98
WSX	1.03	ANH	1.00
WSH	1.04	NES	1.02
ANH	1.07	YKY	1.08
YKY	1.12	WSH	1.09
SRN	1.17	SRN	1.18
NWT	1.17	NWT	1.20

Comments

- Please indicate the units of the explanatory variable, and whether it was expressed in logs.
- Use asterisks to denote significance level: *** (1%), ** (5%) and * (10%)
- P values should be based on cluster robust standard errors
- In the case of random effects please report Stata's output "R2 overall"
- Please use the following naming convention to assign a name to each model: company acronym, level of aggregation, model number (eg for Anglian Water's wholesale water model number 1: ANHWW1). Please refer to the table below for company acronyms and level of aggregation acronyms.

Company acronyms	Level of aggregation acronyms
Anglian Water: ANH Hafren Dyfrdwy: HDD Northumbrian Water: NES	**Wholesale water** Treated water distribution: TWD Water resources plus: WRP

Southern Water: SRN Severn Trent England: SVE South West Water: SWB Thames Water: TMS United Utilities: UUW Dŵr Cymru: WSH Wessex Water: WSX Yorkshire Water: YKY Affinity Water: AFW Bristol Water: BRL Portsmouth Water: PRT SES Water: SES South East Water: SEW South Staffs Water: SSC	Water network plus: WWNP Wholesale water: WW **Wholesale wastewater** Sewage collection: SWC Sewage treatment: STW Bioresources: BR Wastewater network plus: WWWNP Bioresources plus: BRP **Residential retail** Bad debt related costs: RDC Other costs: ROC Total costs: RTC

Template for submission of econometric models for consultation (Wholesale Wastewater Network Plus (SWC & SWT))

Econometric model formula:

1. TMSWWWNP1: ln(WWWNP botex plus Network Reinforcement and Reduced Sewer flooding Growth $_{it}$) = $\alpha + \beta_1$ ln(Load$_{it}$) + β_2 (PCT NH3 below 3mg$_{it}$) + β_3 (PCT Bands 1-3 $_{it}$) + ε_{it}

2. TMSWWWNP2: ln(WWWNP botex plus Network Reinforcement and Reduced Sewer flooding Growth $_{it}$) = $\alpha + \beta_1$ ln(Load$_{it}$) + β_2 (PCT NH3 below 3mg$_{it}$) + β_3 (PCT Bands 6 $_{it}$) + ε_{it}

3. TMSWWWNP3: ln(WWWNP botex plus Network Reinforcement and Reduced Sewer flooding Growth $_{it}$) = $\alpha + \beta_1$ ln(Load$_{it}$) + β_2 (PCT NH3 below 3mg$_{it}$) + β_3 Ln(Pumping Capacity per Length $_{it}$) + ε_{it}

4. TMSWWWNP4: ln(WWWNP botex plus Network Reinforcement and Reduced Sewer flooding Growth $_{it}$) = $\alpha + \beta_1$ ln(Load$_{it}$) + β_2 (PCT NH3 below 3mg$_{it}$) + β_3 (PCT Bands 1-3 $_{it}$) + β_4 Ln(Pumping Capacity per Length $_{it}$) + ε_{it}

5. TMSWWWNP5: ln(WWWNP botex plus Network Reinforcement and Reduced Sewer flooding Growth $_{it}$) = $\alpha + \beta_1$ ln(Load$_{it}$) + β_2 (PCT NH3 below 3mg$_{it}$) + β_3 (PCT Bands 6 $_{it}$) + β_4 Ln(Pumping Capacity per Length $_{it}$) + ε_{it}

6. TMSWWWNP6: ln(WWWNP botex plus Network Reinforcement and Reduced Sewer flooding Growth $_{it}$) = $\alpha + \beta_1$ ln(Load$_{it}$) + β_2 (PCT NH3 below 3mg$_{it}$) + β_3 Ln(Pumping Capacity per Length $_{it}$) + β_4 Ln(Urban Rainfall LAD $_{it}$) + ε_{it}

7. TMSWWWNP7: ln(WWWNP botex plus Network Reinforcement and Reduced Sewer flooding Growth $_{it}$) = $\alpha + \beta_1$ ln(Load$_{it}$) + β_2 (PCT NH3 below 3mg$_{it}$) + β_3 Ln(Pumping Capacity per Length $_{it}$) + β_4 Ln(Urban Rainfall MOSA $_{it}$) + ε_{it}

8. TMSWWWNP8: ln(WWWNP botex plus Network Reinforcement and Reduced Sewer flooding Growth $_{it}$) = $\alpha + \beta_1$ ln(Load$_{it}$) + β_2 (PCT NH3 below 3mg$_{it}$) + β_3 Ln(Pumping Capacity per Length $_{it}$) + β_4 Ln(Annual Rainfall$_{it}$) + ε_{it}

9. TMSWWWNP9: ln(WWWNP botex plus Network Reinforcement and Reduced Sewer flooding Growth $_{it}$) = $\alpha + \beta_1$ ln(Load$_{it}$) + β_2 (PCT NH3 below 3mg$_{it}$) + β_3 Ln(Pumping Capacity per Length $_{it}$) + β_4 Ln(Urban Rainfall LAD $_{it}$) + β_5 (PCT Bands 1-3 $_{it}$) + ε_{it}

10. TMSWWWNP10: ln(WWWNP botex plus Network Reinforcement and Reduced Sewer flooding Growth $_{it}$) = $\alpha + \beta_1$ ln(Load$_{it}$) + β_2 (PCT NH3 below 3mg$_{it}$) + β_3 Ln(Pumping Capacity per Length $_{it}$) + β_4 Ln(Urban Rainfall LAD $_{it}$) + β_5 (PCT Bands 6 $_{it}$) + ε_{it}

11. TMSWWWNP11: ln(WWWNP botex plus Network Reinforcement and Reduced Sewer flooding Growth $_{it}$) = $\alpha + \beta_1$ ln(Load$_{it}$) + β_2 (PCT NH3 below 3mg$_{it}$) + β_3 Ln(Pumping Capacity per Length $_{it}$) + β_4 Ln(Urban Rainfall MOSA $_{it}$) + β_5 (PCT Bands 1-3 $_{it}$) + ε_{it}

12. TMSWWWNP12: ln(WWWNP botex plus Network Reinforcement and Reduced Sewer flooding Growth $_{it}$) = $\alpha + \beta_1$ ln(Load$_{it}$) + β_2 (PCT NH3 below 3mg$_{it}$) + β_3 Ln(Pumping Capacity per Length $_{it}$) + β_4 Ln(Urban Rainfall MOSA $_{it}$) + β_5 (PCT Bands 6 $_{it}$) + ε_{it}

13. TMSWWWNP13: ln(WWWNP botex plus Network Reinforcement and Reduced Sewer flooding Growth $_{it}$) = $\alpha + \beta_1$ ln(Load$_{it}$) + β_2 (PCT NH3 below 3mg$_{it}$) + β_3 Ln(Pumping Capacity per Length $_{it}$) + β_4 Ln(Annual Rainfall $_{it}$) + β_5 (PCT Bands 1-3 $_{it}$) + ε_{it}

14. TMSWWWNP14: ln(WWWNP botex plus Network Reinforcement and Reduced Sewer flooding Growth $_{it}$) = $\alpha + \beta_1$ ln(Load$_{it}$) + β_2 (PCT NH3 below 3mg$_{it}$) + β_3 Ln(Pumping Capacity per Length $_{it}$) + β_4 Ln(Annual Rainfall $_{it}$) + β_5 (PCT Bands 6 $_{it}$) + ε_{it}

Description of the dependent variable

All the econometric models presented in this template use the same definition of **Botex Plus for Wholesale Wastewater Network Plus (Collection and Treatment)** as defined by Ofwat in the Stata code below (e.g., see below **botexplusnpww**):

Sewage Collection Botex

```
g botexswc = BM402SC + BM836SC + BM431SC +
BM140SC + BM839ISC + BM839NISC + BM839OSC
+ BC30945SC + CS00036SC + S3024SC + BN4012_SWC - W3032NPSC -
W3036NPSC - APP28RR_WW0002 - APP28RR_WW0003 - B0201DSWWADJ
```

Sewage Treatment Botex

```
g botexswt = BM502ST + BM836ST + BM531ST
+ BM140ST + BM839IST + BM839NIST + BM839OST
+ BC30945ST + CS00036ST + S3024ST + BN4012_SWT - W3032NPST - W3036NPST
- BN5000 + B0312CRO_SWT + B0318NRO_SWT + B0321PRO_SWT + B0324RSO_SWT +
B0327UVO_SWT
```

Sewage Collection Botex Plus

```
g botexplusswc = botexswc + S3023SC + B0337RFO_TOT + B0200DSISWCWWC +
B0200DSISWCWWO
```

Sewage Treatment Botex Plus

```
g botexplusswt = botexswt + S3023ST
```

Wholesale Wastewater Network Plus Botex

```
g botexnpww = botexswc + botexswt
```

Botex Plus Wholesale Wastewater Network Plus (Botex Plus Network Reinforcement and Reduced Sewer flooding Growth lines for SWC and SWT):

```
g botexplusnpww = botexplusswc + botexplusswt
```

Description of the explanatory variables

- **Ln(Load) = Natural Log or Ln(Load) = Ln(STWDP125_21);** kg BOD5/day
- **Ln(Properties) = Natural Log or Ln(Properties) = Ln((BN1178) * 1000);** Number of Properties
- **Length of Sewage: sewerlength** = BN13535_21 + BN13528; Km
- **Ln(Pumping Capacity per Length)** = Ln(S4029 / sewerlength); (kW/Km)
- **PCT NH3 below 3mg** = ((STWDA121) / (load)) * 100; %
- **PCT bands 1-3** = ((STWDP005_21 + STWDP019_21 + STWDP033_21) / (load)) * 100; %
- **PCT bands 6** = ((STWDP105_21) / (load)) * 100; %
- **Ln(Urban Rainfall LAD)** = Ln(urban_rainfall_lad)
- **Ln(Urban Rainfall MOSA)** = Ln(urban_ urban_rainfall_mosa)
- **Ln(Annual Rainfall)** = Ln(annual_rainfall)

Brief comment on the models

- All the models proposed for WWWNP follow our response to question 8 at the "*Assessing Base Cost at PR24*" consultation. The models presented in this template are a continuation of the insights proposed in the consultation response, in particular with the use of Total Load as the main scale driver. We explore a CSV variable, but Load still provided a better performance for the models overall.
- All the models are run using the period of 11 years, 2011-12 to 2021-22.
- Our proposed models provide an R^2 that ranges between [0.907 – 0.963] depending on the model specification.
- The models are reasonably robust to the tests proposed in the guidance. For example, removing years or most/least efficient companies from the sample does not result in changes to the sign of coefficients on the cost drivers.
- We present the results in a way that shows how the models evolve as additional cost drivers are added in the specification.
- Our first two models TMSWWWNP1-2 are the base of the models proposed. They include a scale driver (Load) and the Percentage of treated load at Bands 1 to 3 and 6. The R^2 of these models is around 0.91. In these models, the Bands are not statistically significant.
- However, the next set of models (TMSWWWNP3-5) include Pumping Capacity per Main. The coefficient on this driver is statistically significant and improves the R^2 of the previous two models to around 0.95.
- The next three models (TMSWWWNP6-8) are an extension of model TMSWWWNP3, with different approaches to capturing rainfall drivers. These three approaches all result in statistically significant drivers, increasing the overall performance of the models through the R^2. The results suggest that urban rainfall is a strong driver linked to base costs. This might be a reflection that most of the sewerage undertakers, concerned with the collection, treatment, and safe disposal of sewage, comes from 'urban' rainfall that directly influences operational base costs. Reference to total annual rainfall alone could lead to poor representation of the effect of rainfall on assets especially given the significant spatial variation of rainfall across the industry. Overall, the effect of rainfall is relevant for base costs TMSWWWNP as also illustrated in our SWC models, providing consistency across different levels of aggregation.
- The last set of models TMSWWWNP9-14 are an extension of models TMSWWWNP4 and TMSWWWNP5. These models improve the R^2 (to around 0.96) compared to the previous ones. All drivers are statistically significant versus the previous models where some cost drivers were not showing a statistical effect. This result suggests that the inclusion of urban rainfall and pumping capacity per main to the first two base models TMSWWWNP1 and 2, improve significantly the overall performance of a potential TMSWWWNP models. We believe that models TMSWWWNP9-14 are the more complete ones as candidates to be considered as a new set of WWWNP models.

	TMSWWWNP1	TMSWWWNP2	TMSWWWNP3	TMSWWWNP4
Dependent Variable	Botex + Network Reinforcement & Reduced Sewer flooding Growth	Botex + Network Reinforcement & Reduced Sewer flooding Growth	Botex + Network Reinforcement & Reduced Sewer flooding Growth	Botex + Network Reinforcement & Reduced Sewer flooding Growth
Ln(Load)	0.632*** 0	0.609*** 0	0.646*** 0	0.727*** 0
PCT NH3 below 3mg	0.006*** 0	0.006*** 0	0.005*** 0	0.005*** 0
PCT Bands 1-3	0.023 0.292			0.023* 0.073
PCT Bands 6		-0.005 0.314		
Ln(Pumping Capacity per Length)			0.367*** 0	0.380*** 0
Ln(Urban Rainfall LAD)				
Ln(Urban Rainfall MOSA)				
Ln(Annual Rainfall)				
Constant	-2.749** 0.024	-2.016*** 0.006	-2.984*** 0	-4.106*** 0
Estimation Method (OLS or RE)	RE	RE	RE	RE
N (sample size)	110	110	110	110
Model robustness tests				
R2 adjusted	0.910	0.907	0.947	0.952
RESET test	0.168	0.238	0.572	0.478
VIF (max) (OLS)	5.337	4.349	4.169	5.396
Pooling / Chow test (OLS)	1.000	1.000	0.978	0.992

Normality of model residuals (OLS)	0.002	0.025	0.435	0.044
Heteroskedasticity of model residuals (OLS)	0.278	0.118	0.515	0.603
Test of pooled OLS versus Random Effects (LM test)	0.000	0.000	0.000	0.000
Efficiency Score Distribution	Min: 0.89	Min: 0.87	Min: 0.92	Min: 0.91
	Max: 1.44	Max: 1.42	Max: 1.07	Max: 1.08
Sensitivity of estimated coefficients to removal of most and least efficient company	A	A	G	A
Sensitivity of estimated coefficients to removal of first and last year of the sample	G	G	G	G

	TMSWWWNP5	**TMSWWWNP6**	**TMSWWWNP7**	**TMSWWWNP8**
Dependent Variable	Botex + Network Reinforcement & Reduced Sewer flooding Growth	Botex + Network Reinforcement & Reduced Sewer flooding Growth	Botex + Network Reinforcement & Reduced Sewer flooding Growth	Botex + Network Reinforcement & Reduced Sewer flooding Growth
Ln(Load)	0.691*** 0	0.589*** 0	0.590*** 0	0.675*** 0
PCT NH3 below 3mg	0.005*** 0	0.005*** 0	0.005*** 0	0.005*** 0
PCT Bands 1-3				
PCT Bands 6	-0.004 0.167			
Ln(Pumping Capacity per Length)	0.370*** 0.001	0.375*** 0	0.376*** 0	0.383*** 0
Ln(Urban Rainfall LAD)		0.088*** 0.005		
Ln(Urban Rainfall MOSA)			0.086*** 0.005	
Ln(Annual Rainfall)				0.084*** 0.009

Constant	-3.228*** 0	-2.919*** 0	-2.929*** 0	-3.923*** 0
Estimation Method (OLS or RE)	RE	RE	RE	RE
N (sample size)	110	110	110	110
Model robustness tests				
R2 adjusted	0.949	0.956	0.955	0.954
RESET test	0.677	0.080	0.195	0.070
VIF (max) (OLS)	4.453	6.654	6.636	4.532
Pooling / Chow test (OLS)	0.997	0.971	0.980	0.932
Normality of model residuals (OLS)	0.101	0.744	0.770	0.631
Heteroskedasticity of model residuals (OLS)	0.762	0.106	0.087	0.425
Test of pooled OLS versus Random Effects (LM test)	0.000	0.000	0.000	0.000
Efficiency Score Distribution	Min: 0.93	Min: 0.95	Min: 0.95	Min: 0.93
	Max: 1.09	Max: 1.09	Max: 1.09	Max: 1.09
Sensitivity of estimated coefficients to removal of most and least efficient company	A	A	A	A
Sensitivity of estimated coefficients to removal of first and last year of the sample	A	A	A	A

	TMSWWWNP9	TMSWWWNP10	TMSWWWNP11	TMSWWWNP12

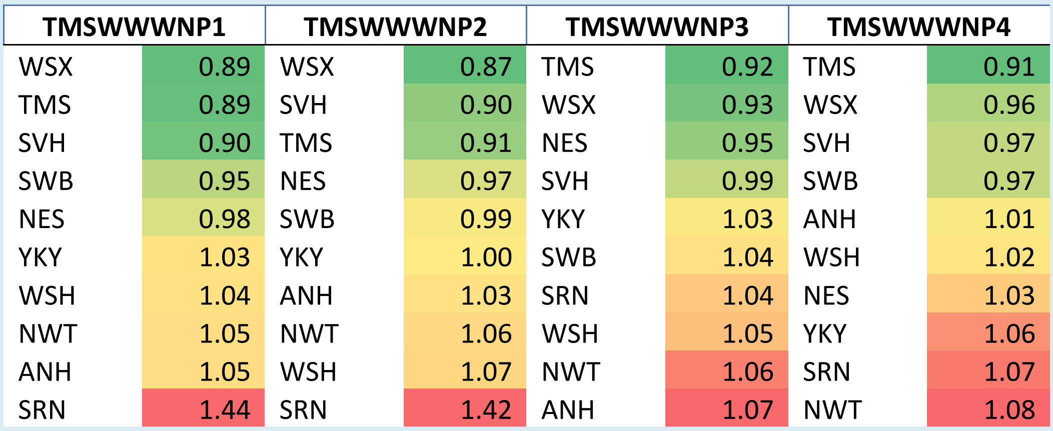
Dependent Variable	Botex + Network Reinforcement & Reduced Sewer flooding Growth	Botex + Network Reinforcement & Reduced Sewer flooding Growth	Botex + Network Reinforcement & Reduced Sewer flooding Growth	Botex + Network Reinforcement & Reduced Sewer flooding Growth
Ln(Load)	0.662*** 0	0.629*** 0	0.664*** 0	0.632*** 0
PCT NH3 below 3mg	0.005*** 0	0.005*** 0	0.005*** 0	0.005*** 0
PCT Bands 1-3	0.022** 0.012		0.022** 0.017	
PCT Bands 6		-0.004* 0.059		-0.004* 0.076
Ln(Pumping Capacity per Length)	0.389*** 0	0.384*** 0	0.391*** 0	0.384*** 0
Ln(Urban Rainfall LAD)	0.095*** 0.003	0.097*** 0.004		
Ln(Urban Rainfall MOSA)			0.091*** 0.003	0.092*** 0.004
Ln(Annual Rainfall)				
Constant	-3.980*** 0	-3.171*** 0	-3.990*** 0	-3.179*** 0
Estimation Method (OLS or RE)	RE	RE	RE	RE
N (sample size)	110	110	110	110
Model robustness tests				
R2 adjusted	0.962	0.960	0.961	0.959
RESET test	0.131	0.036	0.262	0.027
VIF (max) (OLS)	7.843	6.707	7.841	6.704
Pooling / Chow test (OLS)	0.947	0.943	0.968	0.971
Normality of model residuals (OLS)	0.214	0.264	0.216	0.298
Heteroskedasticity of model residuals (OLS)	0.405	0.162	0.345	0.128
Test of pooled OLS versus Random Effects (LM test)	0.000	0.000	0.000	0.000
Efficiency Score Distribution	Min: 0.95	Min: 0.96	Min: 0.94	Min: 0.95

	Max: 1.07	Max: 1.06	Max: 1.07	Max: 1.06
Sensitivity of estimated coefficients to removal of most and least efficient company	A	A	A	A
Sensitivity of estimated coefficients to removal of first and last year of the sample	A	A	A	A

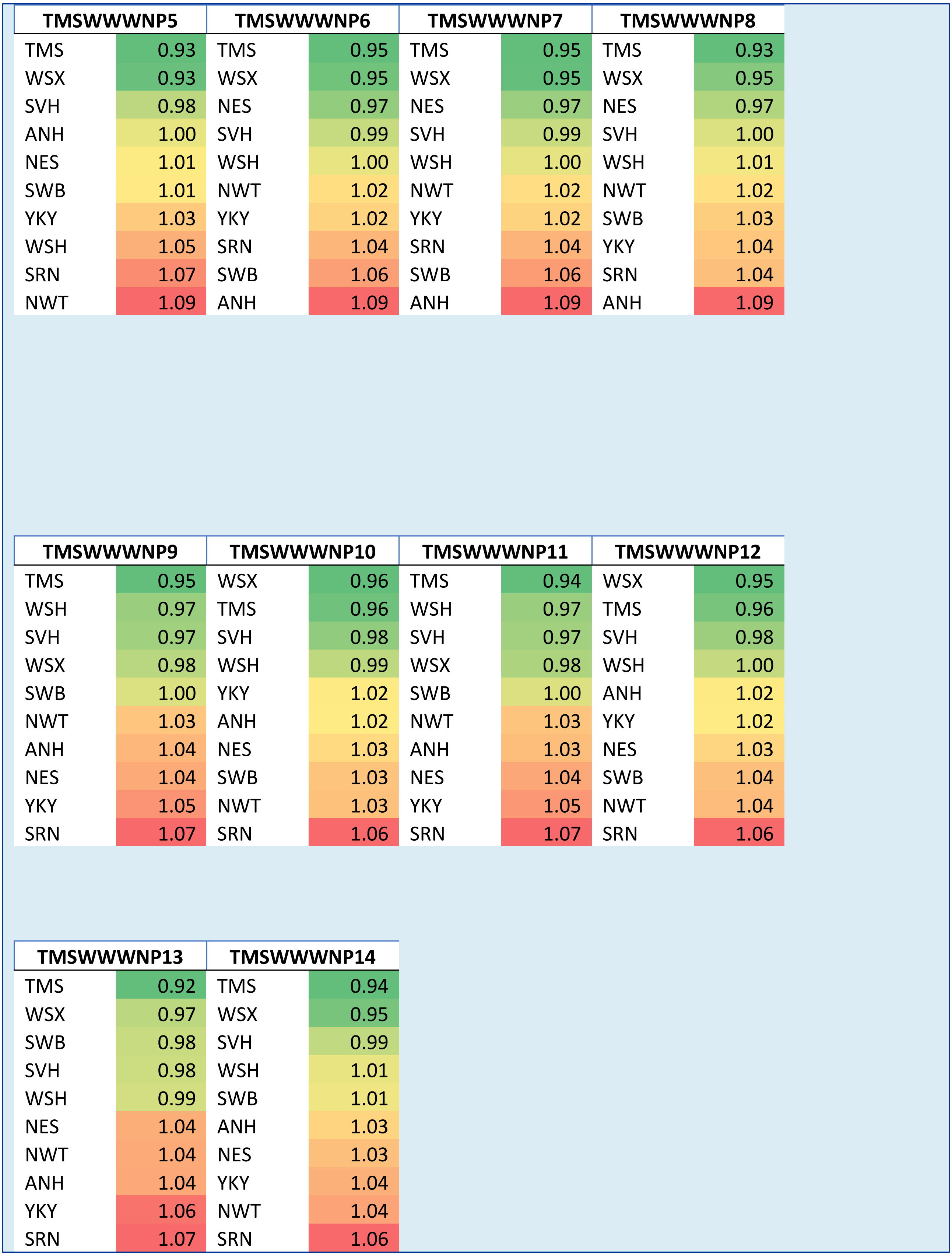
	TMSWWWNP13	**TMSWWWNP14**
Dependent Variable	Botex + Network Reinforcement & Reduced Sewer flooding Growth	Botex + Network Reinforcement & Reduced Sewer flooding Growth
Ln(Load)	0.746*** 0	0.724*** 0
PCT NH3 below 3mg	0.005*** 0	0.005*** 0
PCT Bands 1-3	0.020** 0.038	
PCT Bands 6		-0.004* 0.076
Ln(Pumping Capacity per Length)	0.395*** 0	0.392*** 0
Ln(Urban Rainfall LAD)		
Ln(Urban Rainfall MOSA)		
Ln(Annual Rainfall)	0.082** 0.012	0.092*** 0.008

Constant	-4.889*** 0	-4.269*** 0
Estimation Method (OLS or RE)	RE	RE
N (sample size)	110	110
Model robustness tests		
R2 adjusted	0.958	0.957
RESET test	0.114	0.103
VIF (max) (OLS)	5.749	5.050
Pooling / Chow test (OLS)	0.950	0.865
Normality of model residuals (OLS)	0.338	0.373
Heteroskedasticity of model residuals (OLS)	0.893	0.409
Test of pooled OLS versus Random Effects (LM test)	0.000	0.000
Efficiency Score Distribution	Min: 0.92	Min: 0.94
	Max: 1.07	Max: 1.07
Sensitivity of estimated coefficients to removal of most and least efficient company	A	A
Sensitivity of estimated coefficients to removal of first and last year of the sample	A	A

Efficiency scores distribution



TMSWWWNP1		TMSWWWNP2		TMSWWWNP3		TMSWWWNP4	
WSX	0.89	WSX	0.87	TMS	0.92	TMS	0.91
TMS	0.89	SVH	0.90	WSX	0.93	WSX	0.96
SVH	0.90	TMS	0.91	NES	0.95	SVH	0.97
SWB	0.95	NES	0.97	SVH	0.99	SWB	0.97
NES	0.98	SWB	0.99	YKY	1.03	ANH	1.01
YKY	1.03	YKY	1.00	SWB	1.04	WSH	1.02
WSH	1.04	ANH	1.03	SRN	1.04	NES	1.03
NWT	1.05	NWT	1.06	WSH	1.05	YKY	1.06
ANH	1.05	WSH	1.07	NWT	1.06	SRN	1.07
SRN	1.44	SRN	1.42	ANH	1.07	NWT	1.08



TMSWWWNP5		TMSWWWNP6		TMSWWWNP7		TMSWWWNP8	
TMS	0.93	TMS	0.95	TMS	0.95	TMS	0.93
WSX	0.93	WSX	0.95	WSX	0.95	WSX	0.95
SVH	0.98	NES	0.97	NES	0.97	NES	0.97
ANH	1.00	SVH	0.99	SVH	0.99	SVH	1.00
NES	1.01	WSH	1.00	WSH	1.00	WSH	1.01
SWB	1.01	NWT	1.02	NWT	1.02	NWT	1.02
YKY	1.03	YKY	1.02	YKY	1.02	SWB	1.03
WSH	1.05	SRN	1.04	SRN	1.04	YKY	1.04
SRN	1.07	SWB	1.06	SWB	1.06	SRN	1.04
NWT	1.09	ANH	1.09	ANH	1.09	ANH	1.09

TMSWWWNP9		TMSWWWNP10		TMSWWWNP11		TMSWWWNP12	
TMS	0.95	WSX	0.96	TMS	0.94	WSX	0.95
WSH	0.97	TMS	0.96	WSH	0.97	TMS	0.96
SVH	0.97	SVH	0.98	SVH	0.97	SVH	0.98
WSX	0.98	WSH	0.99	WSX	0.98	WSH	1.00
SWB	1.00	YKY	1.02	SWB	1.00	ANH	1.02
NWT	1.03	ANH	1.02	NWT	1.03	YKY	1.02
ANH	1.04	NES	1.03	ANH	1.03	NES	1.03
NES	1.04	SWB	1.03	NES	1.04	SWB	1.04
YKY	1.05	NWT	1.03	YKY	1.05	NWT	1.04
SRN	1.07	SRN	1.06	SRN	1.07	SRN	1.06

TMSWWWNP13		TMSWWWNP14	
TMS	0.92	TMS	0.94
WSX	0.97	WSX	0.95
SWB	0.98	SVH	0.99
SVH	0.98	WSH	1.01
WSH	0.99	SWB	1.01
NES	1.04	ANH	1.03
NWT	1.04	NES	1.03
ANH	1.04	YKY	1.04
YKY	1.06	NWT	1.04
SRN	1.07	SRN	1.06

Comments

- Please indicate the units of the explanatory variable, and whether it was expressed in logs.
- Use asterisks to denote significance level: *** (1%), ** (5%) and * (10%)
- P values should be based on cluster robust standard errors
- In the case of random effects please report Stata's output "R2 overall"
- Please use the following naming convention to assign a name to each model: company acronym, level of aggregation, model number (eg for Anglian Water's wholesale water model number 1: ANHWW1). Please refer to the table below for company acronyms and level of aggregation acronyms.

Company acronyms	Level of aggregation acronyms
Anglian Water: ANH Hafren Dyfrdwy: HDD Northumbrian Water: NES Southern Water: SRN Severn Trent England: SVE South West Water: SWB Thames Water: TMS United Utilities: UUW Dŵr Cymru: WSH Wessex Water: WSX Yorkshire Water: YKY Affinity Water: AFW Bristol Water: BRL Portsmouth Water: PRT SES Water: SES South East Water: SEW South Staffs Water: SSC	**Wholesale water** Treated water distribution: TWD Water resources plus: WRP Water network plus: WWNP Wholesale water: WW **Wholesale wastewater** Sewage collection: SWC Sewage treatment: STW Bioresources: BR Wastewater network plus: WWWNP Bioresources plus: BRP **Residential retail** Bad debt related costs: RDC Other costs: ROC Total costs: RTC

Template for submission of econometric models for consultation

Econometric model formula:

1. TMSBR1: ln(BR Botex plus Growth Enhancement Capex Per Sludge Produced$_{it}$) = α + β$_1$ ln(Sludge Produced$_{it}$) + β$_2$ (% Load treated in band sizes 1-3$_{it}$) + β$_3$ ln(weighted average density LAD$_{it}$) + ε$_{it}$

2. TMSBR2: ln(BR Botex plus Growth Enhancement Capex Per Sludge Produced$_{it}$) = α + β$_1$ ln(Sludge Produced$_{it}$) + β$_2$ ln(Sewage treatment works per connected property$_{it}$) + ε$_{it}$

3. TMSBR3: ln(BR Botex plus Growth Enhancement Capex Per Sludge Produced$_{it}$) = α + β$_1$ ln(Sludge Produced$_{it}$) + β$_2$ ln(weighted average density LAD$_{it}$) + β$_3$ ln(Work Done in Sludge Disposal by Truck$_{it}$) + ε$_{it}$

4. TMSBR4: ln(BR Botex plus Growth Enhancement Capex Per Sludge Produced$_{it}$) = α + β$_1$ ln(Sludge Produced$_{it}$) + β$_2$ ln(weighted average density LAD$_{it}$) + β$_3$ (ln(weighted average density LAD$_{it}$))^2 + β$_4$ ln(Work Done in Sludge Disposal by Truck$_{it}$) + ε$_{it}$

5. TMSBR5: ln(BR Botex plus Growth Enhancement Capex Per Sludge Produced$_{it}$) = α + β$_1$ ln(Sludge Produced$_{it}$) + β$_2$ ln(Sewage treatment works per connected property$_{it}$) + β$_3$ ln(Work Done in Sludge Disposal by Truck$_{it}$) + ε$_{it}$

6. TMSBR6: ln(BR Botex plus Growth Enhancement Capex Per Sludge Produced$_{it}$) = α + β$_1$ ln(Sludge Produced$_{it}$) + β$_2$ (% Load treated in band sizes6$_{it}$) + β$_3$ ln(weighted average density MSOA$_{it}$) + β$_4$ (ln(weighted average density MSOA$_{it}$))^2 + ε$_{it}$

7. TMSBR7: ln(BR Botex plus Growth Enhancement Capex Per Sludge Produced$_{it}$) = α + β$_1$ ln(Sludge Produced$_{it}$) + β$_2$ (% Load treated in band sizes6$_{it}$) + β$_3$ ln(weighted average density MSOA$_{it}$) + β$_4$ (ln(weighted average density MSOA$_{it}$))^2 + β$_5$ ln(Work Done in Sludge Disposal by Truck$_{it}$) + ε$_{it}$

8. TMSBR8: ln(BR Botex plus Growth Enhancement Capex Per Sludge Produced$_{it}$) = α + β$_1$ ln(Sludge Produced$_{it}$) + β$_2$ (% Load treated in band sizes 1-3$_{it}$) + β$_3$ (%Sludge Disposal Rate$_{it}$) + β$_4$ ln(Work Done in Sludge Disposal by Truck$_{it}$) + ε$_{it}$

9. TMSBR9: ln(BR Botex plus Growth Enhancement Capex Per Sludge Produced$_{it}$) = α + β$_1$ ln(Sludge Produced$_{it}$) + β$_2$ ln(weighted average density LAD$_{it}$) + β$_3$ (ln(weighted average density LAD$_{it}$))^2 + β$_4$ (%Sludge Disposal Rate$_{it}$) + β$_5$ ln(Work Done in Sludge Disposal by Truck$_{it}$) + ε$_{it}$

Description of the dependent variable

All the models use the same definition of **Bioresources Base Costs and Growth Enhancement Capex** as defined by Ofwat under the "Partially Reformed Approach (Option 2)" and as reported in the published PR24 wholesale wastewater dataset.

g Botexbreh = botexsludgetransport + botexsludgetreatment + botexsludgedisposal + BN5000 + BN4012_BIO – W3032SL – W3036SL + B0343SEO_BIO + BN5009

Where:

g botexsludgetransport = BM602STP + BM836STP + BM631STP+ BM140STP + BM839ISTP + BM839NISTP + BM839OSTP+ BC30945STP + CS00036STP

g botexsludgetreatment = BM602SDT + BM836SDT + BM631SDT + BM140SDT + BM839ISDT + BM839NISDT+ BM839OSDT + BC30945SDT + CS00036SDT

g botexsludgedisposal = BM602SDD + BM836SDD + BM631SDD + BM140SDD + BM839ISDD + BM839NISDD + BM839OSDD+ BC30945SDD + CS00036SDD

Description of the explanatory variables

- Ln(Sludge_Produced) = Natural Log or Ln(MP05611)
- %Load Treated in Band Sizes 1-3 = ((STWD012_21 + STWD026_21 + STWD040_21) / (STWD128)) * 100
- %Load Treated in Band Sizes 6 = ((STWD108_21)) / (STWD128)) * 100
- Ln(Weighted Average Density LAD) = Natural Log or Ln(BN4008)
- (Ln(Weighted Average Density LAD))^2 = Natural Log or (Ln(BN4008))^2
- Ln(Weighted Average Density MSOA Area) = Natural Log or Ln(BN4007)
- Ln(Weighted Average Density MSOA Area))^2 = Natural Log or (Ln(BN4007))^2
- Ln(Sewage treatment works per connected property) = Ln(STWC115_21 / BN1178)
- %Sludge Disposal Rate = (BN1621 / MP05611) * 100
- Ln(Work Done in Sludge Disposal by Truck) = Natural Log or Ln(BN1646)

Brief comment on the models

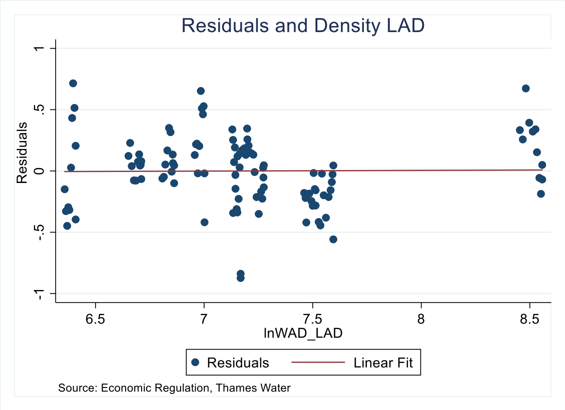
- All models estimated are average cost models i.e. the bioresources base cost and enhancement capex per thousands of tonnes of dried solid sludge produced.
- All models use the time period 2011-12 to 2021-22 and the industry structure is the same as specified by Ofwat.
- Our main insights in the development of the Bioresources models rest on the treatment of *density* and the inclusion of a cost driver that captures the *sludge disposal transportation costs*. For density, we propose including its squared term and for sludge disposal transportation cost, we propose a new driver which is *Total Measure of Work done in sludge disposal operations by truck* in the

models. In these average cost models, we also control for the scale effect of total sludge produced to allow for economies (or diseconomies) of scale.

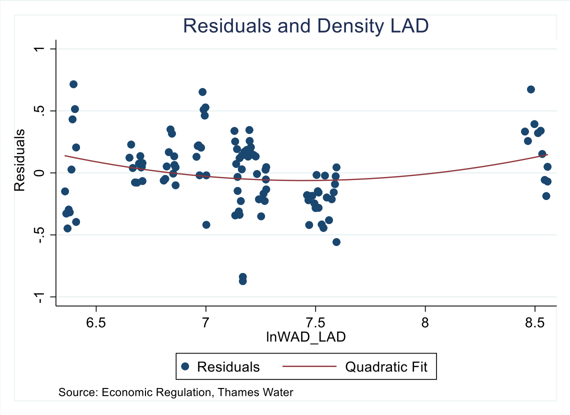
- The effect of the squared term of density adds information in explaining the variation of bioresources costs across the industry with a significant statistical effect. In addition, the inclusion of *Total Measure of Work done in sludge disposal operations by truck* improves the R^2 and fit of the model when compared to the PR19 models and these models pass all necessary tests.
- Adding just the squared term of density to the PR19 BR1 model increases the R^2 from 25% to 29% and improves the dispersion of the industry efficiency scores from 0.67-1.43 to 0.67-1.34. On the other hand, adding just the *Total Measure of Work done in sludge disposal operations by truck* to PR19 BR1 model increases the R^2 from 25% to 40% (see model TMSBR3).
- Including these two drivers in the PR19 BR1 model sees the R^2 increase from 25% to 47%, suggesting a significant improvement in the explanation of the Bioresources costs (see model TMSBR4).
- In addition to the improvement in the R^2, the addition of the new drivers also improves the performance of the current drivers in terms of their signs and statistical significance. Specifically, in the current PR19 BR2 model, the *Number of STWs per property connected* is positive but not significant. Including *Total Measure of Work done in sludge disposal operations by truck* to this model, the *Number of STWs per property connected* becomes statistically significant therefore improving the performance of this driver and the model.
- Also, in the PR19 models BR1 and BR2, the scale driver (log of sludge produced) is positive, indicating diseconomies of scale, and it is insignificant. Including the *Total Measure of Work done in sludge disposal operations by truck*, the scale driver becomes negative although insignificant, indicating the presence of economies of scale.
- However, when we include both the squared density term and the *Total Measure of Work done in sludge disposal operations by truck* to the PR19 BR1 model as seen in models TMSBR4 and TMSBR9, **the scale driver becomes negative and significant as opposed to the PR19 model.** While the inclusion of these drivers changes the sign of the scale variable and therefore moves to an economies of scale interpretation, we note that the guidance states this should not influence model selection. However, the improvements in statistical significance of the scale variable and the explanatory power of the model seen in its R^2 are notable**. Moreover,** this model passes all the statistical specification tests.

- **<u>DENSITY:</u>**
- The Ofwat PR19 BR1 model include the log of the Weighted Average Density using the LAD measure (WAD_LAD) but not its square terms. This assumes that a percentage change in density results in the same percentage change in bioresources costs for all companies. The **negative coefficient of density** in the PR19 models indicate that bioresources unit costs decrease with density i.e. there is the presence of **economies of density.** We disagree with this and we argue that the relationship between bioresources costs and density may be **non-linear** as companies in areas with higher density will not only produce/receive more sludge but might face higher costs in terms of

transportation of dry sludge to landbanks or landfills. We propose including the **squared term** of this variable in the Bioresources models. So, including the **squared term of** density can capture these increased costs as a result of increasing density, showing the importance in distinguishing the heterogeneity across the different levels of density faced by wastewater companies. Plotting the residuals from the PR19 BR1 model against the WAD_LAD variable shows that a linear relationship might not exist between these variables and a non-linear relationship should be explored (see charts below).



Source: Economic Regulation, Thames Water



Source: Economic Regulation, Thames Water

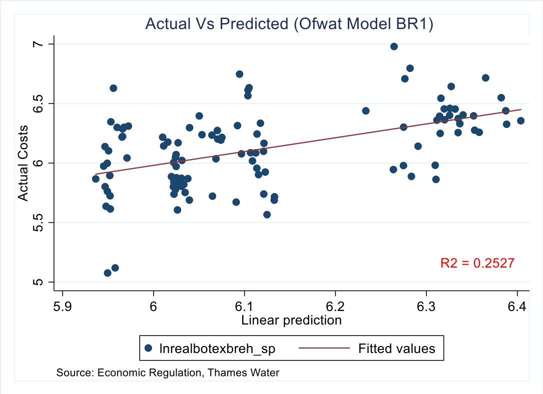
- We include the squared term of WAD_LAD in the PR19 BR1 model and this increases the R^2 from **25% to 29%.** However, we do not include this model in this template as it does not pass the RESET test. Although this model **does not pass the RESET test,** the overall fit of the model is improved as can be seen in the higher R^2. Based on this, we explore including this driver in the models, in addition to the *Total Measure of Work done in sludge disposal operations by truck*. We discuss more about this in the next section.

- **<u>SLUDGE DISPOSAL TRANSPORTATION COSTS.</u>**

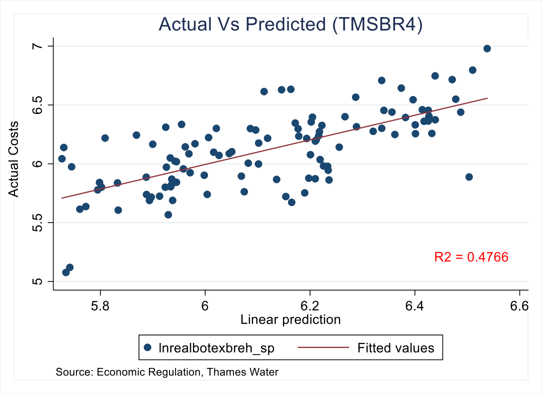
- For PR24, the Bioresources costs models incorporates all elements of the sludge treatment, transport and storage but excluding quality. We feel that the costs associated with **the transportation of the end product (i.e. dried sludge)** to its **final location e.g. landbanks or landfills** has not been adequately captured.

- Wastewater companies operate and face different levels of density according to their geographical location with greater volumes of Biosolids being produced in dense urban areas compared to rural ones. As a result of this heterogeneity in the levels of density, location transportation costs have a significant contribution to the operational costs and we feel that a cost driver that reflects distance covered in the transportation of dry sludge should be controlled for in the Bioresources models. We propose a driver that proxy the distance covered by companies using "*Total Measure of Work done in sludge disposal operations by truck*" which basically uses the accumulated product between sludge mass (in ttds) multiplied by distance travelled (in Kms).

- Our proposed econometric models TMSBR3 and TMSBR5 suggest a significant improvement in explaining the variation of bioresources totex unit costs (without quality) across the industry with a significant statistical effect of this driver on bioresources costs. The robustness of the statistical tests of the models are also strong.
- Including this driver in the two PR19 models i.e. BR1 and BR2, impact the R^2. Specifically, compared to BR1, including the driver in TMSBR3 increases the R^2 by **15 percentage points** and compared to BR2, including the driver increases the R^2 by **20 percentage points,** and both models (i.e. TMSBR3 and TMSBR5) also pass all necessary tests.
- In line with the previous section, there is a positive correlation (0.70) between the sludge disposal transportation driver and the density driver. This is expected as areas with high levels of density might have less availability of landfills and landbanks and so wastewater companies operating in such areas might have to travel a further distance and/or on slower and more costly routes to transport their dried sludge. So, we include the sludge disposal transportation driver and the **squared term of density** mentioned in the previous section. Including these drivers to the PR19 BR1 models as seen in TMSBR4 model increases the R^2 by **22 percentage points** and this model also passes all relevant tests. Also, the overall fit of the model is improved as seen in the charts below.

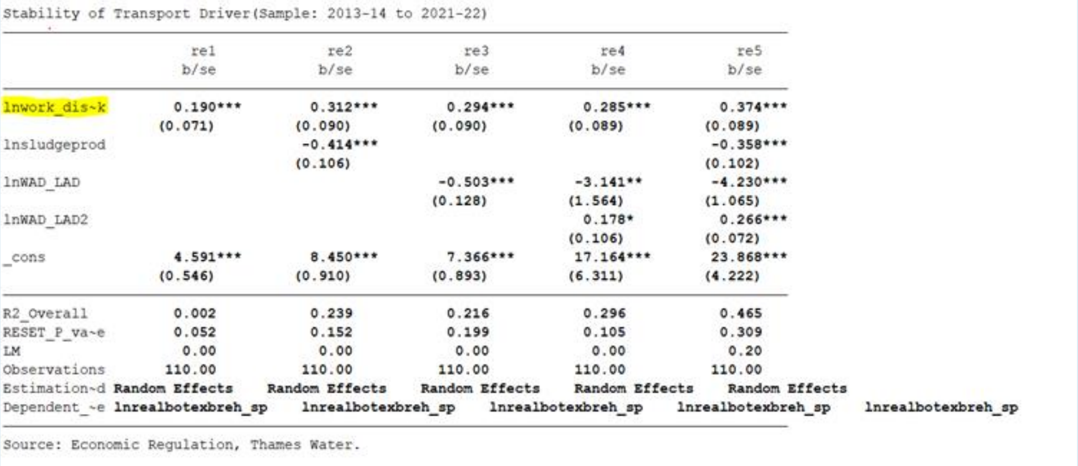


Source: Economic Regulation, Thames Water



Source: Economic Regulation, Thames Water

- This degree of correlation might be a suspicious sign of multicollinearity. However, we explored the potential impact that this driver has when it is included in the models and the stability of the sign of the coefficient as well as its magnitude. We did not see any suggestion of significant changes in the econometric models, as shown in the table below, therefore assuaging concerns of multicollinearity.



Stability of Transport Driver(Sample: 2013-14 to 2021-22)

	re1 b/se	re2 b/se	re3 b/se	re4 b/se	re5 b/se
lnwork_dis~k	0.190***	0.312***	0.294***	0.285***	0.374***
	(0.071)	(0.090)	(0.090)	(0.089)	(0.089)
lnsludgeprod		-0.414***			-0.358***
		(0.106)			(0.102)
lnWAD_LAD			-0.503***	-3.141**	-4.230***
			(0.128)	(1.564)	(1.065)
lnWAD_LAD2				0.178*	0.266***
				(0.106)	(0.072)
_cons	4.591***	8.450***	7.366***	17.164***	23.868***
	(0.546)	(0.910)	(0.893)	(6.311)	(4.222)
R2_Overall	0.002	0.239	0.216	0.296	0.465
RESET_P_va~e	0.052	0.152	0.199	0.105	0.309
LM	0.00	0.00	0.00	0.00	0.20
Observations	110.00	110.00	110.00	110.00	110.00
Estimation~d	Random Effects	Random Effects	Random Effects	Random Effects	Random Effects
Dependent_~e	lnrealbotexbreh_sp	lnrealbotexbreh_sp	lnrealbotexbreh_sp	lnrealbotexbreh_sp	lnrealbotexbreh_sp

Source: Economic Regulation, Thames Water.

- In the table above, we see that the sludge disposal transportation driver exhibits stability in the sign and magnitude. Including the density, squared density and the scale driver (Lnsludgeprod) alternatively does not affect the sign or magnitude of the sludge disposal transportation driver, and all the drivers in the table above have expected signs and magnitude. The negative signs of *Lnsludgeprod* and the density driver indicates **economies of scale and density,** which is expected. Although there is correlation between the drivers, the stability in the results indicates that there might be no presence of multicollinearity.

- **Load Treated in STW:**
- The current PR19 models include the percentage of load treated in small treatment works (bands1-3) as a driver. We also propose using the **percentage of load treated in large treatment works (bands6)** as a driver as this could potentially capture economies of scale since using larger sewage treatment works for larger population centres can be associated with lower unit costs. The **negative sign of this coefficient** confirms this intuition as it portrays the economies of scale that can be achieved when larger sewage treatment works are utilised. Using this variable instead of the loads treated in bands1-3 improves the R^2 by about **3 percentage points** and the model passes all the required tests.
- In addition, with the availability of the **WAD driver at the MSOA level**, which is a smaller unit of measurement compared to the LAD, we explore using this in the model in addition to the **percentage of load treated in large treatment works (bands6)** highlighted in the previous section. A combination of these two drivers can better reflect the **economies of scale and density** of the companies as bioresources costs decrease with density and use of larger treatment works. With this specification as seen in TMSBR6, the R^2 increases by **11 percentage points** compared to the PR19 model BR1 and the model also passes all specification tests.
- Furthermore, we propose the inclusion of *sludge_disposal_rate* driver which to some extent will capture the efficiency of companies in sludge treatment. Ofwat states in the APRs that *"While different technologies exist for sludge treatment, sludge treatment is defined as a technology-neutral service for the purpose of the APR."* This means that other costs incurred by the use of other advanced

technologies, such as digestors, are not captured in the models although wastewater companies incur these costs. Including this driver in addition to the squared density term and the *Total Measure of Work done in sludge disposal operations by truck* driver as seen in models TMSBR8 and TMSBR9, improves the R^2 of the current PR19 BR1 model and also passes all the required tests. Specifically, model TMSBR9 has an R^2 of **51%**, which is twice as large as the **25%** R^2 of the PR19 BR1 model.

	TMSBR1	TMSBR2	TMSBR3	TMSBR4	TMSBR5
Dependent variable	Ln(Bioresources Botex including Enhancement per sludge produced)	Ln(Bioresources Botex including Enhancement per sludge produced)	Ln(Bioresources Botex including Enhancement per sludge produced)	Ln(Bioresources Botex including Enhancement per sludge produced)	Ln(Bioresources Botex including Enhancement per sludge produced)
Ln (Sludge Produced)	0.185 {0.259}	0.17 {0.527}	-0.143 {0.283}	-0.259** {0.018}	-0.139 {0.499}
% load treated in STWs bands 1-3	0.059** {0.036}		0.043** {0.029}	0.026* {0.078}	
% load treated in STWs bands 6					
Ln (Weighted Average Density LAD)	-0.141 {0.296}		-0.263** {0.042}	-3.458*** {0.000}	
Ln (Weighted Average Density LAD)^2				0.216*** {0.000}	
Ln (Weighted Average Density MSOA Area)					
Ln (Weighted Average Density MSOA Area)^2					
% Sludge disposal rate					
Ln (Number of STWs per property)		0.29 {0.173}			0.338* {0.055}
Ln (Work Done in Sludge Disposal by Truck)			0.330*** {0.001}	0.350*** {0.000}	0.326*** {0.000}
Constant	4.745*** {0.007}	6.511*** {0.000}	6.858*** {0.000}	19.856*** {0.000}	7.873*** {0.000}
Estimation method (OLS or RE)	RE	RE	RE	RE	RE
N (sample size)	110	110	110	110	110
Model Robustness					
R2 adjusted	0.253	0.109	0.399	0.477	0.31
RESET test	0.558	0.317	0.658	0.257	0.639

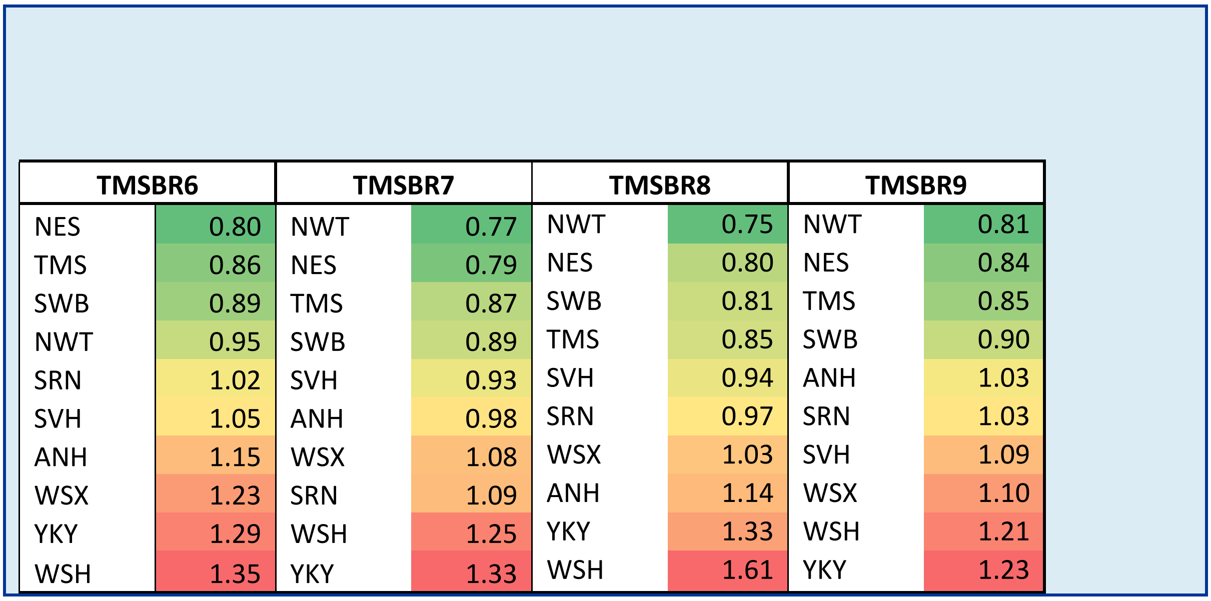
VIF (max)	3.058	3.359	6.856	501.555	5.907
Pooling / Chow test	0.718	0.963	0.777	0.832	0.792
Normality of model residuals	0.453	0.076	0.826	0.432	0.547
Heteroskedasticity of model residuals	0.069	0.464	0.05	0.203	0.189
Test of pooled OLS versus Random Effects (LM test)	0	0	0	0.29	0
Efficiency score distribution (min and max)	Min: 0.67 Max: 1.43	Min: 0.60 Max: 1.46	Min: 0.72 Max: 1.34	Min: 0.77 Max: 1.25	Min: 0.71 Max: 1.42
Sensitivity of estimated coefficients to removal of most and least efficient company	A	A	A	A	A
Sensitivity of estimated coefficients to removal of first and last year of the sample	G	G	A	A	A

	TMSBR6	TMSBR7	TMSBR8	TMSBR9
Dependent variable	Ln(Bioresources Botex including Enhancement per sludge produced)	Ln(Bioresources Botex including Enhancement per sludge produced)	Ln(Bioresources Botex including Enhancement per sludge produced)	Ln(Bioresources Botex including Enhancement per sludge produced)
Ln (Sludge Produced)	0.132 {0.309}	-0.116 {0.409}	-0.211 {0.141}	-0.242*** {0.001}
% load treated in STWs bands 1-3			0.045*** {0.005}	
% load treated in STWs bands 6	-0.020*** {0.001}	-0.006 {0.488}		
Ln (Weighted Average Density LAD)				-4.065*** {0.000}
Ln (Weighted Average Density LAD)^2				0.256*** {0.000}
Ln (Weighted Average Density MSOA Area)	-2.489** {0.011}	-2.724*** {0.005}		
Ln (Weighted Average Density MSOA Area)^2	0.208*** {0.009}	0.199*** {0.010}		
% Sludge disposal rate			0.003 {0.168}	0.005*** {0.000}
Ln (Number of STWs per property)				

Ln (Work Done in Sludge Disposal by Truck)		0.322*** {0.003}	0.277*** {0.001}	0.298*** {0.000}
Constant	13.496*** {0.000}	14.436*** {0.000}	5.963*** {0.000}	22.146*** {0.000}
Estimation method (OLS or RE)	RE	RE	RE	RE
N (sample size)	110	110	110	110
Model Robustness				
R2 adjusted	0.362	0.435	0.365	0.511
RESET test	0.147	0.58	0.751	0.422
VIF (max)	312.28	313.324	7.671	400.229
Pooling / Chow test	0.776	0.833	0.906	0.814
Normality of model residuals	0.089	0.68	0.263	0.706
Heteroskedasticity of model residuals	0.058	0.065	0.086	0.175
Test of pooled OLS versus Random Effects (LM test)	0.043	0.01	0	0.252
Efficiency score distribution (min and max)	Min: 0.73 Max: 1.48	Min: 0.77 Max: 1.33	Min: 0.75 Max: 1.61	Min: 0.81 Max: 1.23
Sensitivity of estimated coefficients to removal of most and least efficient company	A	A	G	G
Sensitivity of estimated coefficients to removal of first and last year of the sample	G	R	G	G

Efficiency scores distribution

TMSBR1		TMSBR2		TMSBR3		TMSBR4		TMSBR5	
NES	0.67	NES	0.60	NWT	0.72	NWT	0.77	NWT	0.71
NWT	0.84	SVH	0.80	NES	0.80	NES	0.80	NES	0.73
SVH	0.87	NWT	0.82	SVH	0.92	TMS	0.86	SVH	0.83
SRN	0.94	SRN	0.97	SWB	0.94	SWB	0.95	SWB	0.94
TMS	0.98	TMS	1.08	ANH	0.95	ANH	0.96	ANH	1.02
SWB	1.03	ANH	1.11	SRN	0.97	SRN	1.02	SRN	1.04
ANH	1.05	SWB	1.12	TMS	1.01	SVH	1.04	TMS	1.05
WSX	1.25	WSX	1.22	WSX	1.09	WSX	1.10	WSX	1.07
YKY	1.32	YKY	1.28	YKY	1.21	WSH	1.15	YKY	1.23
WSH	1.43	WSH	1.46	WSH	1.34	YKY	1.25	WSH	1.42



TMSBR6		TMSBR7		TMSBR8		TMSBR9	
NES	0.80	NWT	0.77	NWT	0.75	NWT	0.81
TMS	0.86	NES	0.79	NES	0.80	NES	0.84
SWB	0.89	TMS	0.87	SWB	0.81	TMS	0.85
NWT	0.95	SWB	0.89	TMS	0.85	SWB	0.90
SRN	1.02	SVH	0.93	SVH	0.94	ANH	1.03
SVH	1.05	ANH	0.98	SRN	0.97	SRN	1.03
ANH	1.15	WSX	1.08	WSX	1.03	SVH	1.09
WSX	1.23	SRN	1.09	ANH	1.14	WSX	1.10
YKY	1.29	WSH	1.25	YKY	1.33	WSH	1.21
WSH	1.35	YKY	1.33	WSH	1.61	YKY	1.23

Company acronyms	Level of aggregation acronyms
Anglian Water: ANH Hafren Dyfrdwy: HDD Northumbrian Water: NES Southern Water: SRN Severn Trent England: SVE South West Water: SWB Thames Water: TMS United Utilities: UUW Dŵr Cymru: WSH Wessex Water: WSX Yorkshire Water: YKY Affinity Water: AFW Bristol Water: BRL Portsmouth Water: PRT SES Water: SES South East Water: SEW South Staffs Water: SSC	Wholesale water Treated water distribution: TWD Water resources plus: WRP Water network plus: WWNP Wholesale water: WW Wholesale wastewater Sewage collection: SWC Sewage treatment: STW Bioresources: BR Wastewater network plus: WWWNP Bioresources plus: BRP Residential retail Bad debt related costs: RDC Other costs: ROC Total costs: RTC

Template for submission of econometric models for consultation

Econometric model formula:

1. TMSRDC1: ln(Bad Debt Related Costs Per HH$_{it}$) = α + β$_1$ ln(Average Bill Size$_{it}$) + β$_2$ (Probability Default$_{it}$) + ε$_{it}$

2. TMSRDC2: ln(Bad Debt Related Costs Per HH$_{it}$) = α + β$_1$ ln(Average Bill Size$_{it}$) + β$_2$ (Income Deprivation$_{it}$) + β$_3$ (Smoothed Transience$_{it}$) + ε$_{it}$

3. TMSRDC3: ln(Bad Debt Related Costs Per HH$_{it}$) = α + β$_1$ ln(Average Bill Size$_{it}$) + β$_2$ ln(Credit_Risk_Score$_{it}$) + β$_3$ (Smoothed_Transience$_{it}$) + ε$_{it}$

4. TMSROC1: ln(Other Costs Per HH$_{it}$) = α + β$_1$ ln(Dual Service Customers$_{it}$) + β$_2$ (Metered Customers$_{it}$) + ε$_{it}$

5. TMSROC2: ln(Other Costs Per HH$_{it}$) = α + β$_1$ ln(Dual Service Customers$_{it}$) + β$_2$ (Metered Customers$_{it}$) + β$_3$ ln(Households_connected$_{it}$) + ε$_{it}$

6. TMSROC3: ln(Other Costs Per HH$_{it}$) = α + β$_1$ ln(Dual Service Customers$_{it}$) + β$_2$ (Metered Customers$_{it}$) + β$_3$ ln(Households_connected$_{it}$) + β$_4$ (Transience$_{it}$) + β$_5$ (Unemployment_LAD$_{it}$) + ε$_{it}$

7. TMSRTC1: ln(Total Costs Per HH$_{it}$) = α + β$_1$ ln(Average Bill Size$_{it}$) + β$_2$ (Probability Default$_{it}$) + β$_3$ (Metered Customers$_{it}$) + ε$_{it}$

8. TMSRTC2: ln(Total Costs Per HH$_{it}$) = α + β$_1$ ln(Average Bill Size$_{it}$) + β$_2$ (Probability Default$_{it}$) + β$_3$ (Metered Customers$_{it}$) + β$_4$ ln(Households connected$_{it}$) + ε$_{it}$

9. TMSRTC3: ln(Total Costs Per HH$_{it}$) = α + β$_1$ ln(Average Bill Size$_{it}$) + β$_2$ (Income Deprivation$_{it}$) + β$_3$ (Smoothed Transience$_{it}$) + β$_4$ (Metered Customers$_{it}$) + β$_5$ ln(Households connected$_{it}$) + ε$_{it}$

10. TMSRTC4: ln(Total Costs Per HH$_{it}$) = α + β$_1$ ln(Average Bill Size$_{it}$) + β$_2$ ln(Credit Risk Score$_{it}$) + β$_3$ (Smoothed Transience$_{it}$) + β$_4$ (Metered Customers$_{it}$) + β$_5$ ln(Households connected$_{it}$) + ε$_{it}$

***These models exclude the financial year 2019-20**

11. TMSRDC4: ln(Bad Debt Related Costs Per HH$_{it}$) = α + β$_1$ ln(Average Bill Size$_{it}$) + β$_2$ (Income Deprivation$_{it}$) + β$_3$ (Transience$_{it}$) + ε$_{it}$

12. TMSRTC5: ln(Total Costs Per HH$_{it}$) = $\alpha + \beta_1$ ln(Average Bill Size$_{it}$) + β_2 (Income Deprivation$_{it}$) + β_3 (Transience$_{it}$) β_4 (Metered Customers$_{it}$) + β_5 ln(Households_connected$_{it}$) + ε_{it}

****These Models Use the Sample Period 2013-14 to 2018-19**

13. TMSRDC5: ln(Bad Debt Related Costs Per HH$_{it}$) = $\alpha + \beta_1$ ln(Average Bill Size$_{it}$) + β_2 (Income Deprivation$_{it}$) + β_3 (Transience$_{it}$) + ε_{it}

14. TMSRTC6: ln(Total Costs Per HH$_{it}$) = $\alpha + \beta_1$ ln(Average Bill Size$_{it}$) + β_2 (Income Deprivation$_{it}$) + β_3 (Transience$_{it}$) + β_4 (Metered Customers$_{it}$) + β_5 ln(Households_connected$_{it}$) + ε_{it}

Description of the dependent variables

All the models use the same definition of **Total Costs, Other Costs and Bad Debt Costs** as defined by Ofwat in the Stata code.

g sTC_tr = BM9030 + BM9002 + BM9003 + BM9007 + BM9033 + BM9019 + BM9020 + s_depreciation + netrecharges

g DC_t = BM9002 + BM9003

g sOC_tr = sTC_tr - DC_t

Description of the explanatory variables

- Ln(Households_connected) = Natural Log or Ln(hh_t) = Ln(R3017 + R3019 + R3021 + R3018 + R3020 + R3022)
- Ln(Average Bill Size) = Natural Log or Ln(rev_hh) = Ln(rev_t * 1000 / hh_t)
- Transience = totalmigration; % total internal and international migration
- Smoothed_Transience = (totalmigration$_{it}$ + totalmigration$_{it-1}$ + totalmigration$_{it-2}$) / 3[1]
- Income Deprivation = incomescore_interpolated: % income deprivation (2015 and 2019 values are interpolated to fill in the missing years 2016-2018) and 2019-20 data is used for the subsequent years up to 2021-22.
- Probability of Default = eq_lpcf62; % of households with default
- Ln(Credit_Risk_Score) = Natural Log or Ln(eq_rgc102); credit risk score derived from all Insight data (Score Range is 000-200)
- Unemployment = unemploymentrate; % of unemployment rate at LAD level. This variable is calculated using ONS unemployment data at LAD level (from Nomis).

[1] There is no material difference when using a three-year or two-year moving average for smoothed transience.

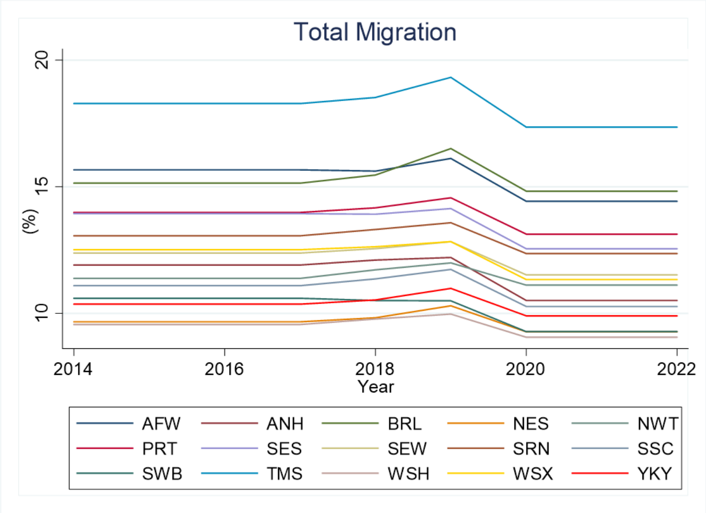
- Metered Customers = hhm_hh: % of metered customers
- Dual Service Customers = hhdu_hh: % of dual service customers

Brief comment on the models

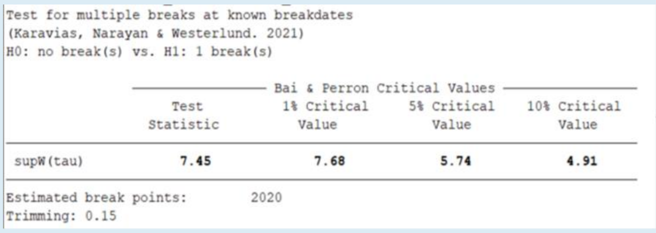
- TMSRDC1 – TMSRTC4 models are run using the period of 9 years, i.e. 2013-14 to 2021-22.
- For Models TMSRDC4 and TMSRTC5, we propose excluding the 2019-20 financial period in order to account for the macro shock associated with that period due to the effect of the Covid pandemic on **transience (please see below).**
- Following from the previous point, we also propose using the time period **2013-14 to 2018-19** in models TMSRDC5 and TMSRTC6. This is also an attempt to mitigate the effect of the pandemic on the estimation.
- Most of the models remain robust with regards to the signs of the coefficients when the **least and most efficient companies** are removed from the sample. However, this is not the case for when the **last and first year** of the sample period are removed. This is also the case with the current retail models from PR19.
- All proposed models satisfy the robustness checks with **high importance status.**

- **<u>TRANSIENCE:</u>**
- We are concerned about the transience variable currently captured by **totalmigration** in the current PR19 models. This variable is expected to be a **positive** driver of costs as it could increase bad debt costs, make debt recovery more difficult or increase account management costs i.e. setting up, closing and transferring accounts. However, in the current PR19 version of the **bad debt models**, **totalmigration** has a **negative** coefficient, which goes against core **economic rationale**.
- Also, in the current PR19 version of the **Total Cost Models, totalmigration** has a **much smaller coefficient (0.004)** when compared to the PR19 models **(0.037).** This is about **9 times smaller** than the previous PR19 model and is correspondingly a source of potential concern.

- We suspect that the observed issues with **totalmigration** in the bad debt and total retail costs models might be related to issues arising from the COVID-19 pandemic.



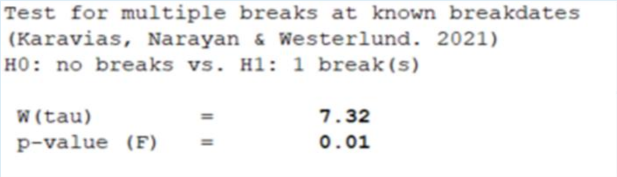
- The chart above shows a macro shock to migration for 2019-2020 which affected all companies. Migration seemed to be on an upward trend which was reversed by the pandemic in 2020 (**Note: ONS migration figures are reported mid-year i.e. up to 30 June and so 2019-2020 will include early pandemic effects on Migration**). Asides from this, there is a structural break in 2015-16 as the ONS changed the methodology used in calculating these figures. So for the periods before 2015-16, 2015-16 figures were used in order to mitigate this.
- We also suspect another structural break occurred in 2019-20 and testing for this using test by Ditzen, J., Karavias, Y. & Westerlund, J. (2021) confirms this.



```
Test for multiple breaks at known breakdates
(Karavias, Narayan & Westerlund. 2021)
H0: no break(s) vs. H1: 1 break(s)

                    ─────────────── Bai & Perron Critical Values ───────────────
                        Test          1% Critical     5% Critical     10% Critical
                      Statistic          Value           Value            Value
──────────────────────────────────────────────────────────────────────────────────
 supW(tau)               7.45             7.68            5.74             4.91
──────────────────────────────────────────────────────────────────────────────────
Estimated break points:        2020
Trimming: 0.15
```

Testing for *unknown* break dates shows 2020 as a structural break date.



```
Test for multiple breaks at known breakdates
(Karavias, Narayan & Westerlund. 2021)
H0: no breaks vs. H1: 1 break(s)

 W(tau)         =         7.32
 p-value  (F)   =         0.01
```

On the other hand, on testing 2020 as a *known* break date, we reject the null hypothesis that there are no breaks in the data

- Ideally, we would like to control for **time effects** in these models through the inclusion of time dummies but we recognise the potential issue when the estimated results need to be taken to forecast cost efficiency allowances.
- To mitigate these issues, we propose:

  i) Excluding **2019-20 period** from the models which include **transience** measured as **totalmigration** as a driver.

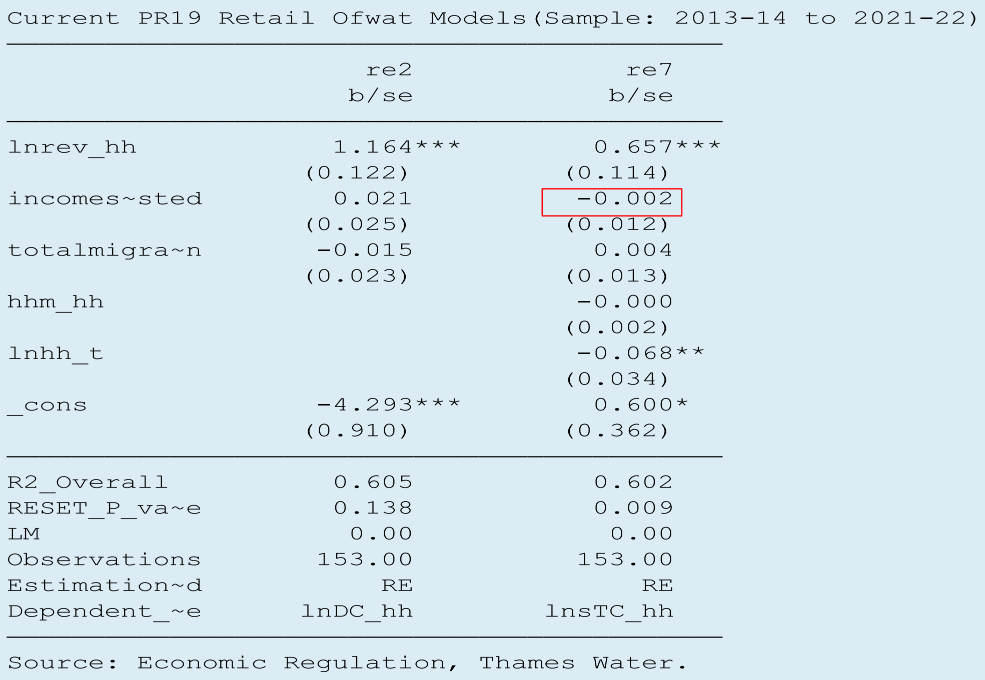
ii) Splitting the sample period and using the period up until **2018-19.**

iii) **Smoothing** the **totalmigration** variable using a 3-year moving average.

- Any of the above propositions yield models whose coefficients follow expected **economic rationale** and magnitudes are within the expected range.

- **INCOME DEPRIVATION:**

- We are also concerned about the **income deprivation driver** captured as **incomescore_unadjusted** in the current version of the PR19 models. The most recent data was released in **2019** and this is assumed to hold for the periods of **2019-20** and beyond. Given the pandemic effect in these years, we are concerned that this variable might not accurately capture the level of deprivation witnessed as a result of the pandemic.
- In the current PR19 version of the **Total Retail Cost models,** this variable has a **negative** effect on cost, which appears counterintuitive to **economic rationale**. It is difficult to argue that greater deprivation should reduce debt costs faced by companies.



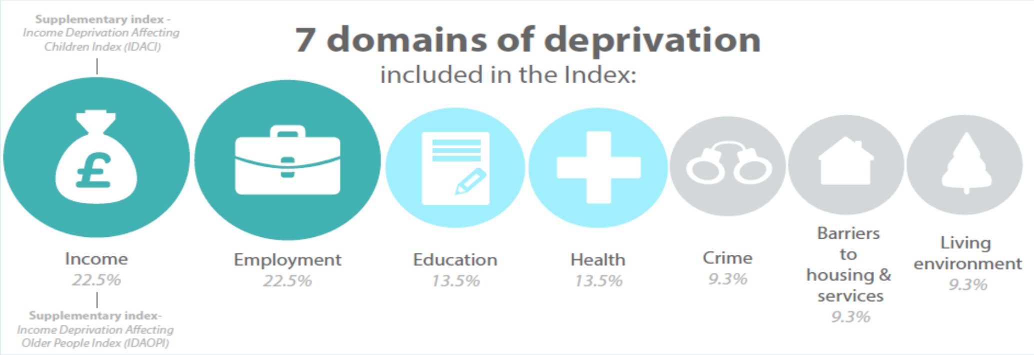
```
Current PR19 Retail Ofwat Models(Sample: 2013-14 to 2021-22)
--------------------------------------------
                         re2             re7
                        b/se            b/se
--------------------------------------------
lnrev_hh               1.164***        0.657***
                     (0.122)         (0.114)
incomes~sted           0.021          -0.002
                     (0.025)         (0.012)
totalmigra~n          -0.015           0.004
                     (0.023)         (0.013)
hhm_hh                                -0.000
                                     (0.002)
lnhh_t                                -0.068**
                                     (0.034)
_cons                 -4.293***        0.600*
                     (0.910)         (0.362)
--------------------------------------------
R2_Overall             0.605           0.602
RESET_P_va~e           0.138           0.009
LM                      0.00            0.00
Observations          153.00          153.00
Estimation~d              RE              RE
Dependent_~e         lnDC_hh         lnsTC_hh
--------------------------------------------
Source: Economic Regulation, Thames Water.
```

- To address this, we propose the following:

i) Use of **Credit Risk Score** to capture **deprivation** instead of the **incomescore_unadjusted.** There is a high **negative correlation** between these two variables (**-0.8662)** and the expectation is that this driver would have a **negative** effect on costs as higher credit scores should indicate less deprivation. The models proposed using this variable are robust, improve the R^2, passes all necessary tests and are consistent with **economic intuition.** Since this data is published yearly, it could be an adequate substitute for the **income score**.

ii) Use of **incomescore_interpolated**. We consider this to be the second best alternative as it mitigates some of the issues with the **incomescore_unadjusted** variable and appears with the right signs in the models.

iii) Combined use of **Credit Risk Score** and **Unemployment rate** to capture deprivation. Employment(Unemployment) and Barriers to Housing and Services (Credit Risk Score) make up at least **30%** of the current **Index of Deprivation,** they can be good measures of deprivation.


Supplementary index - Income Deprivation Affecting Children Index (IDACI)
7 domains of deprivation
included in the Index:
Income 22.5%
Employment 22.5%
Education 13.5%
Health 13.5%
Crime 9.3%
Barriers to housing & services 9.3%
Living environment 9.3%
Supplementary index- Income Deprivation Affecting Older People Index (IDAOPI)

- **ADDITIONAL DRIVERS FOR OTHER RETAIL COSTS:**
- Customer Service Costs make up about 50% of other costs in the industry and potentially has drivers different from the Ofwat's proposed drivers of Other retail Costs, such as **transience** and **deprivation** (measured as unemployment)[2].
- Given that Customer Service Costs make up a material part of Other Costs, the absence of transience and deprivation from the Other Retail Costs model needs to be re-evaluated based on their performance in the model (see TMSROC3).
- Models **TMSROC1 and TMSROC2** are the PR19 version of the **Other Retail Costs Models** with an R^2 of **0.138.** Also, these models are not robust to the removal of the first and last years, and the least and most efficient companies. However, by adding **transience and deprivation** to the model as seen in **TMSROC3,** the R^2 increases to **0.199** (about 2% is explained by transience and about 4% is explained by deprivation). In addition, the **TMSROC3** is robust to the removal of the first and last years, and the least and most efficient companies unlike **TMSROC1 and TMSROC2.** Lastly, **TMSROC3** has a smaller dispersion in the efficiency scores (0.80-1.36) compared to the PR19 models (0.82-1.51).
- Given the above, we propose that **transience and deprivation** should be included in the **Other Retail Costs Models** as they are potential drivers of customer service costs and improve the performance of the Other Retail Costs Models.

[2] This is based on preliminary exploratory Customer Service Models. These models are still being developed and are not ready to be submitted at this time.

	TMSRDC1	TMSRDC2	TMSRDC3
Dependent variable	Bad Debt related Costs Per HH	Bad Debt related Costs Per HH	Bad Debt related Costs Per HH
Ln(Average Bill Size)	1.188*** (0.000)	1.114*** (0.000)	1.210*** (0.000)
Probability of Default	0.024 (0.209)		
Income Deprivation		0.052* (0.057)	
Transience			
Smoothed Transience		0.027 (0.296)	0.024 (0.379)
Ln(Credit Risk Score)			-3.275* (0.074)
Dual Service Customers			
Metered Customers			
Ln(Households Connected)			
Unemployment			
Constant	-4.899*** (0.000)	-4.918*** (0.000)	11.4 (0.195)
Estimation method (OLS or RE)	RE	RE	RE
Model robustness tests			
N (sample size)	153	153	153
R2 adjusted	0.615	0.641	0.621
RESET test	0.092	0.04	0.132
VIF (max)	1.009	1.357	1.113
Pooling / Chow test	0.063	0.13	0.226
Normality of model residuals	0	0	0
Heteroskedasticity of model residuals	0	0	0
Test of pooled OLS versus Random Effects (LM test)	0	0	0
Efficiency score distribution (min and max)	Min: 0.63 Max: 1.82	Min: 0.68 Max: 1.84	Min: 0.63 Max: 1.91
Sensitivity of estimated coefficients to removal of most and least efficient company	G	A	G
Sensitivity of estimated coefficients to removal of first and last year of the sample	R	R	R

	TMSROC1	TMSROC2	TMSROC3	TMSRTC1	TMSRTC2	TMSRTC3
Dependent variable	Other Costs Per HH	Other Costs Per HH	Other Costs Per HH	Total Costs Per HH	Total Costs Per HH	Total Costs Per HH
Ln(Average Bill Size)				0.519*** (0.000)	0.621*** (0.000)	0.651*** (0.000)
Probability of Default				0.011 (0.415)	0.02 (0.141)	
Income Deprivation						0.039** (0.047)
Transience			0.053*** {0.000}			
Smoothed Transience						0.048*** (0.004)
Ln(Credit Risk Score)						
Dual Service Customers	0.002** (0.025)	0.003*** (0.000)	0.007*** {0.000}			
Metered Customers	0.0004371 (0.809)	0.000405 (0.834)	0.004* {0.064}	0.001 (0.668)	0.003 (0.348)	0.005 (0.119)
Ln(Households Connected)		-0.049 (0.117)	-0.172*** {0.000}		-0.081*** (0.005)	-0.125*** (0.001)
Unemployment			0.029 {0.360}			
Constant	2.718*** (0.000)	3.355*** (0.000)	3.992*** {0.000}	0.101 (0.835)	0.344 (0.388)	0.091 (0.797)
Estimation method (OLS or RE)	RE	RE	RE	RE	RE	RE
N (sample size)	153	153	153	153	153	153
Model robustness tests						
R2 adjusted	0.127	0.138	0.199	0.613	0.636	0.624
RESET test	0.586	0.124	0.7	0	0.006	0.146
VIF (max)	1.00	2.119	4.687	2.396	3.186	4.581
Pooling / Chow test	0.96	0.993	0.191	0.858	0.8	0.976
Normality of model residuals	0.264	0.409	0.588	0.038	0.153	0.337
Heteroskedasticity of model residuals	0.683	0.997	0.131	0.589	0.175	0.265
Test of pooled OLS versus Random Effects (LM test)	0	0	0	0	0	0
Efficiency score distribution (min and max)	Min: 0.82 Max: 1.54	Min: 0.82 Max: 1.51	Min: 0.80 Max: 1.36	Min: 0.82 Max: 1.32	Min: 0.78 Max: 1.24	Min: 0.77 Max: 1.22

Sensitivity of estimated coefficients to removal of most and least efficient company	R	R	A	G	G	A
Sensitivity of estimated coefficients to removal of first and last year of the sample	R	R	A	G	A	A

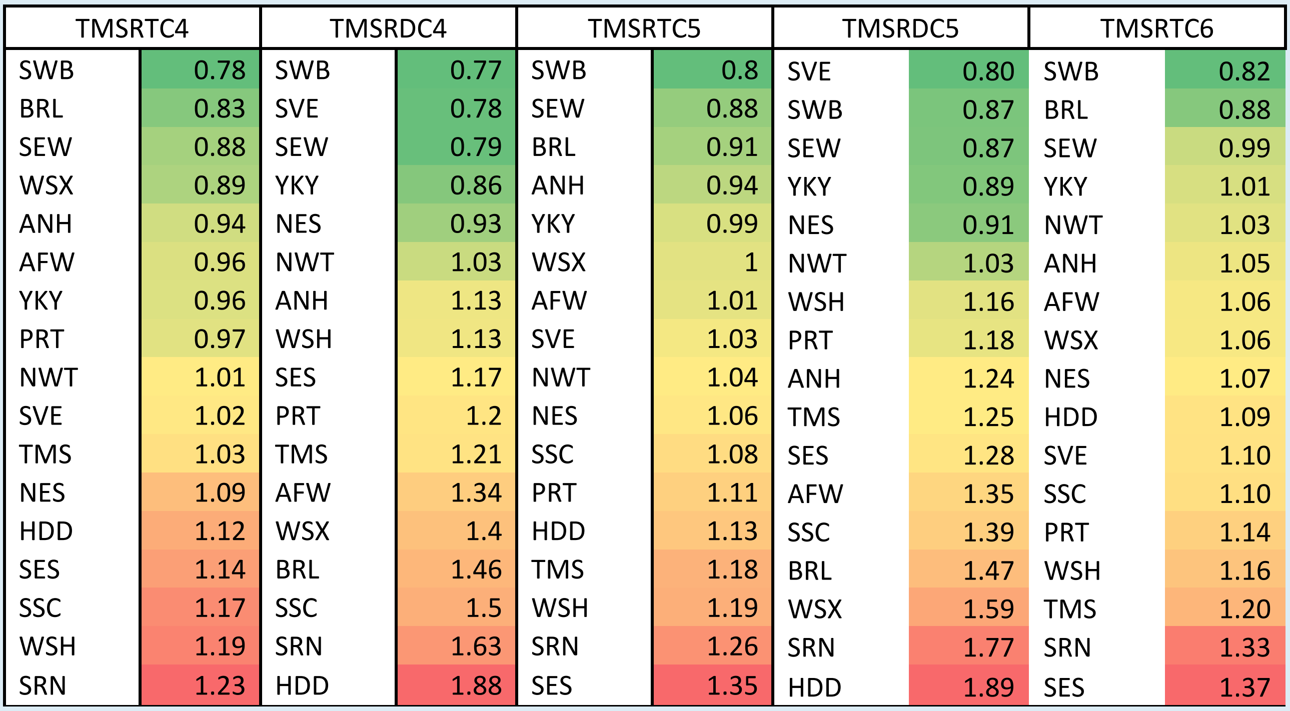
	TMSRTC4	**TMSRDC4**	**TMSRTC5**	**TMSRDC5**	**TMSRTC6**
Dependent variable	Total Costs Per HH	Bad Debt related Costs Per HH	Total Costs Per HH	Bad Debt related Costs Per HH	Total Costs Per HH
Ln(Average Bill Size)	0.799*** (0.000)	1.131*** (0.000)	0.614*** (0.000)	1.022*** (0.000)	0.710*** (0.000)
Probability of Default					
Income Deprivation		0.073** (0.012)	0.051*** (0.006)	0.097*** (0.003)	0.070*** (0.004)
Transience		0.021 (0.338)	0.038*** (0.007)	0.026 (0.181)	0.056** (0.018)
Smoothed Transience	0.050*** (0.000)				
Ln(Credit Risk Score)	-2.727*** (0.000)				
Dual Service Customers					
Metered Customers	0.002 (0.483)		0.005* (0.099)		0.004 (0.180)
Ln(Households Connected)	-0.167*** (0.000)		-0.113*** (0.000)		-0.181*** (0.002)
Unemployment					
Constant	13.943*** (0.000)	-5.255*** (0.000)	0.082 (0.816)	-5.041*** (0.000)	0.051 (0.917)
Estimation method (OLS or RE)	RE	RE	RE	RE	RE
N (sample size)	153	136	136	102	102
Model robustness tests					
R2 adjusted	0.635	0.693	0.683	0.761	0.711
RESET test	0.083	0.007	0.005	0.506	0.688
VIF (max)	4.988	1.333	4.284	1.383	4.427
Pooling / Chow test	0.991	0.99	1	1	1
Normality of model residuals	0.325	0	0.274	0	0.428

Heteroskedasticity of model residuals	0.163	0	0.377	0	0.377
Test of pooled OLS versus Random Effects (LM test)	0	0	0	0	0
Efficiency score distribution (min and max)	Min: 0.78 Max: 1.23	Min: 0.77 Max: 1.88	Min: 0.80 Max: 1.35	Min: 0.80 Max: 1.89	Min: 0.82 Max: 1.37
Sensitivity of estimated coefficients to removal of most and least efficient company	G	G	A	G	G
Sensitivity of estimated coefficients to removal of first and last year of the sample	G	R	A	G	G

Efficiency scores distribution

TMSRDC1		TMSRDC2		TMSRDC3	
SEW	0.63	SEW	0.68	SEW	0.63
SWB	0.7	SWB	0.73	SWB	0.7
SVE	0.79	SVE	0.76	SVE	0.75
YKY	0.86	YKY	0.87	YKY	0.82
SES	0.91	NES	0.96	SES	0.92
ANH	0.96	SES	1	NES	0.96
NES	0.97	ANH	1.05	ANH	0.97
PRT	1.04	PRT	1.06	TMS	1
NWT	1.08	NWT	1.07	PRT	1.02
TMS	1.14	TMS	1.1	NWT	1.04
AFW	1.17	WSH	1.17	AFW	1.15
WSH	1.2	AFW	1.18	WSX	1.15
WSX	1.2	WSX	1.26	WSH	1.21
BRL	1.44	BRL	1.3	BRL	1.29
SRN	1.45	SSC	1.48	SRN	1.41
SSC	1.59	SRN	1.5	SSC	1.6
HDD	1.82	HDD	1.84	HDD	1.91

TMSROC1		TMSROC2		TMSROC3		TMSRTC1		TMSRTC2		TMSRTC3	
YKY	0.38	YKY	0.37	BRL	0.80	SEW	0.82	SWB	0.78	SWB	0.77
SSC	0.41	SSC	0.42	WSX	0.82	SWB	0.83	SEW	0.82	SEW	0.81
PRT	0.43	PRT	0.42	PRT	0.83	ANH	0.83	ANH	0.85	BRL	0.84
SWB	0.45	SWB	0.43	ANH	0.85	SVE	0.9	YKY	0.94	ANH	0.9
TMS	0.45	TMS	0.48	SWB	0.89	WSX	0.92	WSX	0.94	WSX	0.93
WSX	0.49	WSX	0.49	NWT	0.89	YKY	0.93	AFW	0.98	AFW	0.94
NWT	0.5	NWT	0.5	AFW	0.95	AFW	0.93	SVE	0.99	YKY	1
ANH	0.51	ANH	0.52	YKY	0.98	NES	0.99	NES	1	SVE	1.02
SVE	0.53	SVE	0.55	HDD	0.99	PRT	1	BRL	1.01	PRT	1.05
SEW	0.54	SEW	0.57	TMS	1.00	BRL	1.01	NWT	1.02	NWT	1.07
AFW	0.56	BRL	0.57	SEW	1.05	NWT	1.01	PRT	1.03	TMS	1.09
BRL	0.57	SRN	0.6	SVE	1.06	SSC	1.06	SSC	1.08	NES	1.1
SRN	0.58	AFW	0.6	SSC	1.07	TMS	1.12	HDD	1.08	SSC	1.11
NES	0.62	NES	0.63	SRN	1.15	WSH	1.18	WSH	1.16	HDD	1.13
WSH	0.66	WSH	0.64	WSH	1.16	HDD	1.23	TMS	1.23	SRN	1.18
HDD	0.76	HDD	0.69	NES	1.26	SRN	1.24	SRN	1.24	WSH	1.21
SES	0.84	SES	0.82	SES	1.36	SES	1.32	SES	1.25	SES	1.22



TMSRTC4		TMSRDC4		TMSRTC5		TMSRDC5		TMSRTC6	
SWB	0.78	SWB	0.77	SWB	0.8	SVE	0.80	SWB	0.82
BRL	0.83	SVE	0.78	SEW	0.88	SWB	0.87	BRL	0.88
SEW	0.88	SEW	0.79	BRL	0.91	SEW	0.87	SEW	0.99
WSX	0.89	YKY	0.86	ANH	0.94	YKY	0.89	YKY	1.01
ANH	0.94	NES	0.93	YKY	0.99	NES	0.91	NWT	1.03
AFW	0.96	NWT	1.03	WSX	1	NWT	1.03	ANH	1.05
YKY	0.96	ANH	1.13	AFW	1.01	WSH	1.16	AFW	1.06
PRT	0.97	WSH	1.13	SVE	1.03	PRT	1.18	WSX	1.06
NWT	1.01	SES	1.17	NWT	1.04	ANH	1.24	NES	1.07
SVE	1.02	PRT	1.2	NES	1.06	TMS	1.25	HDD	1.09
TMS	1.03	TMS	1.21	SSC	1.08	SES	1.28	SVE	1.10
NES	1.09	AFW	1.34	PRT	1.11	AFW	1.35	SSC	1.10
HDD	1.12	WSX	1.4	HDD	1.13	SSC	1.39	PRT	1.14
SES	1.14	BRL	1.46	TMS	1.18	BRL	1.47	WSH	1.16
SSC	1.17	SSC	1.5	WSH	1.19	WSX	1.59	TMS	1.20
WSH	1.19	SRN	1.63	SRN	1.26	SRN	1.77	SRN	1.33
SRN	1.23	HDD	1.88	SES	1.35	HDD	1.89	SES	1.37

Comments

- Please indicate the units of the explanatory variable, and whether it was expressed in logs.
- Use asterisks to denote significance level: *** (1%), ** (5%) and * (10%)
- P values should be based on cluster robust standard errors
- In the case of random effects please report Stata's output "R2 overall"
- Please use the following naming convention to assign a name to each model: company acronym, level of aggregation, model number (eg for Anglian Water's wholesale water model number 1: ANHWW1). Please refer to the table below for company acronyms and level of aggregation acronyms.

Company acronyms	Level of aggregation acronyms
Anglian Water: ANH Hafren Dyfrdwy: HDD Northumbrian Water: NES Southern Water: SRN Severn Trent England: SVE South West Water: SWB Thames Water: TMS United Utilities: UUW Dŵr Cymru: WSH Wessex Water: WSX Yorkshire Water: YKY Affinity Water: AFW Bristol Water: BRL Portsmouth Water: PRT SES Water: SES South East Water: SEW South Staffs Water: SSC	**Wholesale water** Treated water distribution: TWD Water resources plus: WRP Water network plus: WWNP Wholesale water: WW **Wholesale wastewater** Sewage collection: SWC Sewage treatment: STW Bioresources: BR Wastewater network plus: WWWNP Bioresources plus: BRP **Residential retail** Bad debt related costs: RDC Other costs: ROC Total costs: RTC