South West
Water

Peninsula House, Rydon Lane, Exeter, EX2 7HR
www.southwestwater.co.uk

11 January 2023

Daniel Mitchell
Principal Economist
PR24 and Beyond
Ofwat
Centre City Tower
7 Hill Street
Birmingham
B5 4UA

By email: CostAssessment@ofwat.gov.uk

Dear Daniel

SUBMISSION OF BASE ECONOMETRIC MODELS AHEAD OF THE SPRING 2023 CONSULTATION

This is a brief covering letter for the base econometric models South West Water (and Bristol Water) have proposed ahead of the consultation process.

We have submitted via Sharepoint a suite of completed submission templates, Stata .do files, regression outputs. The calculation/distribution of the efficiency scores is included in the model templates as well as in the Stata regression output for each cost model. The water wholesale dataset and bioresources data sets used are as per the standard .do file published by Ofwat. There are two alternative retail data sets and a wholesale wastewater data set provided with the submission.

We have worked with Oxera on this submission to ensure that our proposed models meet the PR24 base cost assessment principles set out in the guidance. On sharepoint there are separate folders containing the files for each template model area as set out below. In high level summary our models are:

Wholesale water

- SWBTWD1 which is a treated water distribution model that includes a time trend from 2017
- SWBWW1 which is a wholesale water aggregated model that includes a time trend from 2017
- SWBWW2 which is a wholesale water aggregated model including weighted average treatment complexity and a time trend from 2017.

Wholesale wastewater

- SWBSWC1 a sewerage model which includes annual rainfall
- SWBSWC2 a sewerage model which includes population density and % of the sewer asset base after 2001
- SWBSWT1 a sewage treatment model which includes a composite treatment index measure.

Bioresources

- SWBBR1 which includes pipeline intersiting, liming and incumbent treatment as exogenous location factors
- SWBBR2 which uses weighted average size bands to reflect scale.

Retail

- A range of other retail models that includes combinations of metered, number of services, migration, time trends and Covid dummy factors.
- A range of bad debt costs including alternative deprivation and covid dummy factors
- A total cost model with equivalent explanatory variables to the disaggregated models.

We would happy Bristol Water also worked with Wessex Water and Reckon on retail models, which are an additional set of models that we have not submitted to avoid duplication.

Our work with Oxera has been extensive and there are areas we plan to continue to explore further in advance of the modelling consultation. In particular, We have undertaken significant research into the cost implications of seasonality and UV treatment. We have not been able to demonstrate any robust econometric relationships between publicly available and exogenous variables in the time available. Nevertheless, we remain of the view that seasonality and UV treatment are likely to be cost drivers to a significant extent.

Please let me know if you have any questions or would like to discuss any aspects of our model submission further.

Kind regards

Iain McGuffog
Director of Strategy & Regulation
E: iain.mcguffog@bristolwater.co.uk
M: 07976 269968

Template for submission of econometric models for consultation

Econometric model formula:

1. SWBTWD1: ln(botex plus TWD$_{it}$) = α + β_1 (ln(lengthsofmain$_{it}$)) + β_2 (ln(weighted average density LAD$_{it}$)) + β_3 (ln(weighted average density LAD$_{it}$)^2) + β_4 (ln(Average pumping head (distribution)$_{it}$) + β_5 (timetrend2017$_t$) + ε_{it}

2. SWBWW1: ln(botex plus WW$_{it}$) = α + β_1 (ln(properties$_{it}$)) + β_2 (pctwatertreated36$_{it}$) + β_3 (ln(weighted average density LAD$_{it}$)) + β_4 (ln(weighted average density LAD$_{it}$)^2) + β_5 (ln(Average pumping head (distribution)$_{it}$) + β_6 (timetrend2017$_t$) + ε_{it}

3. SWBWW2: ln(botex plus WW$_{it}$) = α + β_1 (ln(properties$_{it}$)) + β_2 (ln(weighted average level of treatment complexity$_{it}$)) + β_3 (ln(weighted average density LAD$_{it}$)) + β_4 (ln(weighted average density LAD$_{it}$)^2) + β_5 (ln(Average pumping head (distribution)$_{it}$) + β_6 (timetrend2017$_t$) + ε_{it}

Description of the dependent variable

- SWBWW1 and SWBWW2 : wholesale water botex including network reinforcement (code: Botex+NR_WW in Interface_real), as reported in the published PR24 wholesale dataset.
- SWBTWD1 : treated water distribution botex including network reinforcement (code: Botex+NR_TWD in Interface_real), as reported in the published PR24 wholesale dataset.

Description of the explanatory variables

- Total properties (code: properties in Interface_real), as reported in the published wholesale dataset.
- Weighted average density LAD (code: WAD – LAD – water in Interface_real), as reported in the published wholesale dataset.
- Weighted average level of treatment complexity (code: wac in Interface_real), as reported in the published wholesale dataset.
- Average pumping head – distribution (code: BN4870 in Stata dataset (nominal)), as reported in the published wholesale dataset.

- Timetrend2017 : time trend starting at 1 in 2016/17 (i.e. 0 up to 2015/16, 1 in 2016/17, 2 in 2017/18, 3 in 2018/19, etc.).
- Lengths of main (code: lengthsofmain in Interface_real), as reported in the published PR24 wholesale dataset.

Brief comment on the models

- we propose to include 'average pumping head (distribution)' as the use of this driver better aligns with operational insight, in that it should reflect differences in power usage and topography across companies;
- unsurprisingly we have therefore found it to be a better cost driver than 'booster pumping stations per length of main' in terms of statistical significance in both treated water distribution and wholesale water models;
- additionally, we propose to also include a time trend variable (beginning in the financial year 2016-17) as this captures the impact of the structural break we have identified in base costs since 2016-17;
- in 2016-17 costs increased for 15 of the 17 companies supplying wholesale water services, with costs for the industry as a whole increasing by 12.1%; this increase in costs has not been reversed, with costs continuing to increase on average in the years 2018-2020. We believe this includes the impact of the service-cost relationship and rising trend in water costs linked to water outcomes and leakage and supply interruption performance;
- the inclusion of both average pumping head and the time trend variable improves the goodness of fit of the treated water distribution model, vis-à-vis Ofwat's PR19 model;
- we also note that both cost drivers perform well against Ofwat's robustness sensitivities, with the drivers' coefficients maintaining significance levels when the first and last years of the sample are removed, and when the most and least efficient companies are removed;
- the full historical data has been used for all proposed models.
- although we have investigated the use of the new seasonality variable in the modelling we have not found its current form (ratio between peak and average demand during the year) to perform particularly well statistically. However we still think seasonal variation in demand is an issue for us and that the current modelling does not captured the higher costs faced during peak demands. We will aim to explore this further in later discussions with Ofwat.

	SWBTWD1	SWBWW1	SWBWW2
Dependent variable	TWD botex including network reinforcement	WW botex including network reinforcement	WW botex including network reinforcement
Ln(Lengths of main)	1.061*** (0.000)	NA	NA
Ln(weighted average density LAD)	-2.958*** (0.000)	-2.115*** (0.000)	-1.990*** (0.000)
Ln(weighted average density LAD) ^{2}	0.227*** (0.000)	0.142*** (0.000)	0.133*** (0.000)
Ln(average pumping head – distribution)	0.282*** (0.000)	0.271** (0.012)	0.267** (0.018)
Timetrend2017	0.022*** (0.000)	0.018** (0.024)	0.018** (0.020)
Ln(properties)	NA	1.062*** (0.000)	1.057*** (0.000)
Water treated in bands 3-6 (%)	NA	0.003* (0.063)	NA
Ln(weighted average level of treatment complexity)	NA	NA	0.220* (0.071)
Constant	2.262 (0.160)	-3.532*** (0.005)	-3.999*** (0.002)
Estimation method (OLS or RE)	RE	RE	RE
N (sample size)	187	187	187
Model robustness tests			
R2 adjusted	0.963	0.971	0.97
RESET test	0.881	0.992	0.982
VIF (max)	203.81	212.572	201.276
Pooling / Chow test	0.998	0.987	0.964
Normality of model residuals	0.321	0.015	0.306
Heteroskedasticity of model residuals	0.245	0	0

Test of pooled OLS versus Random Effects (LM test)	0	0	0
Efficiency score distribution (min and max)	Min: 0.68 Max: 1.27	Min: 0.74 Max: 1.42	Min: 0.72 Max: 1.40
Sensitivity of estimated coefficients to removal of most and least efficient company	G	G	G
Sensitivity of estimated coefficients to removal of first and last year of the sample	G	G	G

Efficiency scores SWBTWD1

Rank	Company	Efficiency score
1	SWB	67.84%
2	SES	88.59%
3	NWT	91.80%
4	PRT	93.61%
5	WSX	93.84%
6	SVE	95.30%
7	SSC	98.61%
8	HDD	100.07%
9	NES	102.76%
10	TMS	106.64%
11	SRN	106.83%
12	SEW	107.70%
13	ANH	112.82%
14	AFW	119.34%
15	WSH	119.47%
16	YKY	125.35%
17	BRL	126.97%

Efficiency scores SWBWW1

Rank	Company	Efficiency score
1	SSC	74.21%
2	PRT	85.62%
3	SWB	91.12%
4	SEW	92.47%
5	SVE	95.48%
6	ANH	98.22%
7	AFW	98.79%
8	WSX	99.51%
9	NES	102.73%
10	NWT	103.64%
11	TMS	105.61%
12	HDD	105.80%
13	YKY	106.27%
14	SES	113.92%
15	BRL	114.53%
16	WSH	121.34%
17	SRN	141.89%

Efficiency scores SWBWW2

Rank	Company	Efficiency score
1	SSC	72.29%
2	PRT	84.20%
3	SWB	91.40%
4	SEW	92.36%
5	ANH	95.76%
6	SVE	97.21%
7	WSX	98.08%
8	AFW	98.62%
9	NES	103.26%
10	NWT	104.01%
11	TMS	106.36%

12	YKY	107.45%
13	HDD	107.92%
14	BRL	112.77%
15	SES	116.17%
16	WSH	120.84%
17	SRN	140.28%

Template for submission of econometric models for consultation

Econometric model formula:

1. SWBSWC1: $\ln(\text{botex plus SWC}_{it}) = \alpha + \beta_1 (\ln(\text{total sewer length}_{it})) + \beta_2 (\ln(\text{pumping capacity/km of sewer}_{it})) + \beta_3 (\ln(\text{number of properties per km of sewer length}_{it})) + \beta_4 (\ln(\text{annual rainfall – wastewater}_{it})) + \varepsilon_{it}$

2. SWBSWC2: $\ln(\text{botex plus SWC}_{it}) = \alpha + \beta_1 (\ln(\text{total sewer length}_{it})) + \beta_2 (\ln(\text{pumping capacity/km of sewer}_{it})) + \beta_3 (\ln(\text{weighted average population density based on LADs, weighted by population}_{it})) + \beta_4 (\ln((\text{weighted average population density based on LADs, weighted by population}_{it})^2)) + \beta_5 (\text{\% of the sewer asset base constructed after 2001}_{it}) + \varepsilon_{it}$

3. SWBSWT1: $\ln(\text{botex plus SWT}_{it}) = \alpha + \beta_1 (\ln(\text{load}_{it})) + \beta_2 (\text{pctbands6}_{it}) + \beta_3 (\ln(\text{CompositeTreatment}_{it}) + \varepsilon_{it}$

Description of the dependent variable

- SWBSWC1 and SWBSWC2 : sewage collection botex plus including network reinforcement and reduced sewer flooding growth lines (code: botex_sc_sewerflood_rein in Interface_real), as reported in the published PR24 wholesale dataset.
- SWBSWT1 : sewage treatment botex including reduced sewer flooding growth lines (code: botex_st_sewerflood in Interface_real), as reported in the published PR24 wholesale dataset.

Description of the explanatory variables

SWBSWC1 and SWBSWC2 :

- Total sewer length (code: sewerlength in Interface_real), as reported in the published PR24 wholesale dataset.
- Pumping capacity/km of sewer (code: pumpingcapperlength in Interface_real), as reported in the published PR24 wholesale dataset.
- Number of properties per km of sewer length (code: properties in Interface_real), as reported in the published PR24 wholesale dataset.

- Annual rainfall – wastewater (code: BN4512 in Stata dataset (nominal)), as reported in the published wholesale dataset.
- Weighted average population density based on LADs, weighted by population (code: WAD_LAD in Interface_real), as reported in the published wholesale dataset.
- % of the sewer asset base constructed after 2001 (code: BB2370 in Stata dataset (nominal) divided by total sewer length above).

SWBSWT1:

- Load (code: Load in Interface_real), as reported in the published PR24 wholesale dataset.
- Pctbands6 : % of load treated at bands 6 and above, defined as STWDP105_21/Load*100
- CompositeTreatment: defined as the sum of IndexPhosphorus, IndexAmmonia and IndexUV, multiplied by 100, with:
  - IndexPhosphorus being the % of load treated with phosphorus consent below 0.5mg/L, i.e. STWDP121_21/Load;
  - IndexAmmonia being the % of load treated with ammonia consent below 3mg/L, as per Ofwat's PR19 modelling, i.e. (STWDA121 + STWDA122_21)/Load;
  - IndexUV being the % of load treated with UV consent, irrespective of the threshold of 30mW/s/cm2, i.e. (STWDU026+STWDU025)/Load.

Brief comment on the models

- The higher costs incurred to treat UV is not reflected in Ofwat's PR19 models. Therefore our view is that PR24 models should account for UV cost specificities, either by including standalone UV measures or in the form of a composite treatment variable. We have proposed one model below with a composite measure aiming to aggregate phosphorus, ammonia and UV treatment as a single measure. This is still a work in progress and we will continue to develop models that reflect the higher costs associated with UV treatment.
- As there is not a direct relationship between UV treatment costs and the dose level, we have not retained a specific threshold for UV consent and have included both the treatment below and above 30mW/s/cm2.
- The composite treatment measure performs particularly well statistically as it is always significant at the 1% level in both the proposed model and its different sensitivities.
- The coefficient of the variable 'pctbands6' is also statistically significant at the 10% level (very close from being significant at the 5% level) and its sign is in line with operational insights, i.e. reflecting economies of scale and the fact that larger bands

tend to have lower unit costs. We also note that this cost driver is 'only' significant at the 10% level in Ofwat's modelling, with a similar coefficient of -0.011 and a p-value of 0.053.

- the model fit is similar to Ofwat's comparable SWT2 model with a R2 of 0.84 vs 0.85
- the min-max range is also similar (from 0.67 initially to 0.68).
- the full historical data has been used for all proposed models.

SWBSCW1 and SWBSWC2:

- On the sewage collection models, we propose to include both 'annual rainfall' and the '% of the asset base constructed after 2001' as drivers of sewage collection costs, as these align with operational insight and perform well in the models;
- in years and areas where rainfall is higher, the costs of collecting sewage will be higher, due to the higher volume of sewage that must be collected and transported;
- companies with newer asset bases will face lower capital maintenance costs, as these assets are newer and generally in a better state of repair;
- both drivers are highly significant and perform well across Ofwat's robustness sensitivities when removing the most and least efficient companies, and the first and last years of the sample;
- both models present a better goodness of fit than Ofwat's updated PR19 models;
- the spread of efficiency scores is similar, or actually slightly tighter, than Ofwat's updated PR19 models;
- the full historical data has been used for all proposed models.

	SWBSWC1	SWBSWC2	SWBSWT1
Dependent variable	SWC botex plus	SWC botex plus	SWT botex plus
Ln(Total sewer length)	0.850*** (0.000)	0.804*** (0.000)	NA
Ln(pumping capacity/km of sewer)	0.364*** (0.002)	0.535*** (0.000)	NA
Ln(Number of properties per km of sewer length)	1.034*** (0.000)	NA	NA
Ln(Annual rainfall – wastewater)	0.151*** (0.000)	NA	NA
Ln(Weighted average population density based on LADs, weighted by population)	NA	-2.182** (0.016)	NA
Ln(Weighted average population density based on LADs, weighted by population)[2]	NA	0.165*** (0.005)	NA

% of the sewer asset base constructed after 2001	NA	-0.014** (0.017)	NA
Ln(Load)	NA	NA	0.891*** (0.000)
pctbands6	NA	NA	-0.013* (0.056)
Ln(CompositeTreatment)	NA	NA	0.171*** (0.000)
Constant	-9.459*** (0.000)	3.031 (0.414)	-6.146*** (0.000)
Estimation method (OLS or RE)	RE	RE	RE
N (sample size)	110	110	110
R2 adjusted	0.928	0.919	0.843
RESET test	0.344	0	0.134
VIF (max)	2.725	402.881	2.541
Pooling / Chow test	0.859	0.821	0.999
Normality of model residuals	0.012	0.092	0.089
Heteroskedasticity of model residuals	0.306	0.202	0.999
Test of pooled OLS versus Random Effects (LM test)	0	0	0
Efficiency score distribution (min and max)	Min: 0.93 Max: 1.14	Min: 0.89 Max: 1.19	Min : 0.88 Max: 1.57
Sensitivity of estimated coefficients to removal of most and least efficient company	G	G	G
Sensitivity of estimated coefficients to removal of first and last year of the sample	G	G	G

Efficiency scores SWBSWC1

Rank	Company	Efficiency score
1	WSX	93.31%
2	SVH	94.69%
3	NES	95.44%
4	SRN	96.05%
5	WSH	97.94%
6	NWT	98.14%
7	ANH	98.88%
8	YKY	103.45%
9	SWB	108.70%
10	TMS	113.78%

Efficiency scores SWBSWC2

Rank	Company	Efficiency score
1	ANH	89.01%
2	SWB	94.27%
3	NES	98.44%
4	TMS	99.03%
5	WSX	100.54%
6	SRN	102.42%
7	NWT	103.91%
8	SVH	106.35%
9	WSH	108.19%
10	YKY	118.98%

Efficiency scores SWBSWT1

Rank	Company	Efficiency score
1	SVH	88.38%
2	TMS	89.23%
3	NES	92.86%
4	WSX	93.05%
5	SWB	95.56%
6	ANH	98.32%
7	WSH	111.21%
8	YKY	112.55%
9	NWT	113.26%
10	SRN	156.86%

Template for submission of econometric models for consultation

Econometric model formula:

1. SWBBR1: ln(botex bioresources$_{it}$) = α + β_1 (ln(total sludge produced$_{it}$)) + β_2 (ln(weighted average population density based on LADs, weighted by population$_{it}$)) + β_3 (% of intersiting work done by pipeline$_{it}$) + β_4 (% of sludge treated by the incumbent using raw liming$_{it}$) + β_5 (% of sludge treated by the incumbent in total$_{it}$)) + ε_{it}

2. SWBBR2: ln(botex plus SWC$_{it}$) = α + β_1 (ln(total sludge produced$_{it}$)) + β_2 (Weighted average sewage treatment work size band$_{it}$)) + ε_{it}

Description of the dependent variable

- SWBBR1 and SWBBR2 : botex bioresources including bioresources quality enhancement opex (codes: botex_bio in Interface_real + B0343SEO_BIO in Stata dataset (real)), as reported in the published PR24 wholesale dataset.

Description of the explanatory variables

- Total sludge produced (code: sludgeprod in Interface_real), as reported in the published PR24 wholesale dataset.
- Weighted average population density based on LADs, weighted by population (code: WAD_LAD in Interface_real), as reported in the published PR24 wholesale dataset.
- % of intersiting work undertaken by pipeline (codes: BN1640 divided by BN1643, all multiplied by 100, in Stata dataset (nominal)), as reported in the published wholesale dataset.
- % of sludge treated by the incumbent using raw liming (code: BN5612INC_21, multiplied by 100, in Stata dataset (nominal)), as reported in the published wholesale dataset.
- % of sludge treated by the incumbent (code: BN5619INC_21 multiplied by 100, in Stata dataset (nominal)), as reported in the published wholesale dataset.
- The weighted average sewage treatment work size band (WASB, created by multiplying the size band number (1 to 6) by the proportion of load treated at the respective size band (codes: SWTD012_21, STWD026_21, STWD040_21, STWD054_21, STWD068_21, and STWD108_21 in Stata dataset (nominal) for load

treated at size bands 1 to >5 respectively, using STWDP125_21 for the total load treated), as reported in the published wholesale dataset.

Brief comment on the models

- the full historical data has been used for all proposed models;
- the coefficients are highly significant in both models and are robust to the removal of both the first and last years and most and least efficient companies;
- both models present an improvement in the fit of the model against Ofwat's updated PR19 models, with the R-squared increasing;
- model SWBBR1 captures the impact on costs of transporting and treating sludge by different methods;
- model SWBBR2 captures some of the density impact previously captured by the 'sewage treatment works per property' variable, while also capturing the proportion of load treated at different size bands (through the weighting used in the measure) which replaces the impact previously captured by 'the percentage of sewage treated in size bands 1-3';
- the signs of the coefficients are intuitive and align with operational insight.

	SWBBR1	SWBBR2
Dependent variable	Botex bioresources	Botex bioresources
Ln(Total sludge produced)	1.258*** (0.000)	1.130*** (0.000)
Ln(Weighted average population density based on LADs, weighted by population)	-0.286** (0.012)	NA
% of intersiting work done by pipeline	-0.009*** (0.000)	NA
% of sludge treated by the incumbent using raw liming	0.009*** (0.000)	NA
% of sludge treated by the incumbent	-0.034*** (0.000)	NA

Weighted average sewage treatment work size band	NA	-1.156*** (0.000)
Constant	3.335*** (0.000)	5.029*** (0.000)
Estimation method (OLS or RE)	RE	RE
N (sample size)	110	110
R2 adjusted	0.867	0.835
RESET test	0.449	0.83
VIF (max)	3.652	2.206
Pooling / Chow test	0.727	0.186
Normality of model residuals	0.049	0.026
Heteroskedasticity of model residuals	0.051	0.076
Test of pooled OLS versus Random Effects (LM test)	0.115	0.009
Efficiency score distribution (min and max)	Min: 0.74 Max: 1.48	Min: 0.73 Max: 1.58
Sensitivity of estimated coefficients to removal of most and least efficient company	A	G
Sensitivity of estimated coefficients to removal of first and last year of the sample	G	G

Efficiency scores SWBBR1

Rank	Company	Efficiency score
1	NES	74.19%
2	SWB	88.69%
3	SVH	92.69%
4	SRN	94.20%
5	NWT	108.01%
6	TMS	108.78%
7	ANH	111.68%
8	YKY	111.90%
9	WSX	126.75%
10	WSH	147.98%

Efficiency scores SWBBR2

Rank	Company	Efficiency score
1	NES	72.52%
2	SVH	91.87%
3	SRN	92.39%
4	NWT	96.29%
5	SWB	97.25%
6	ANH	102.78%
7	TMS	104.21%
8	WSX	108.50%
9	YKY	126.87%
10	WSH	157.78%

Template for submission of econometric models for consultation

Econometric model formula:

<u>Other Retail Costs</u>

1. ln(sOC_hh$_{it}$)= α + β1.(% of dual service connections$_{it}$) + β2(ln(Total households connected$_{it}$))+ β3(% of wwater only service connections$_{it}$) + β4 (% of metered connections$_{it}$) + β5(totalmigration$_{it}$) + β6(timetrend$_{t}$) + β7(covid_2020$_{t}$) + ε_{it}

2. ln(sOC_ss$_{it}$)= α + β1(% of metered services$_{it}$) + β2(% of services that are wastewater$_{it}$) + β3(ln(total number of connected services$_{it}$)) + β4(% of totalmigration$_{it}$) + + β5(covid_2020$_{t}$) +β6(timetrend$_{t}$) + ε_{it}
**Using a 2.0 weighting (explained later in the document)*

3. . ln(sOC_ss$_{it}$)= α + β1(% of metered services$_{it}$) + β2(% of services that are wastewater$_{it}$) + β3(ln(total number of connected services$_{it}$)) + β4(% of totalmigration$_{it}$) + + β5(covid_2020$_{t}$) +β6(timetrend$_{t}$) + ε_{it}
**Using a 1.3 weighting (explained later in the document)*

<u>Bad Debt Costs</u>

4. ln(DC_hh$_{it}$)= α + β1(ln (average bill size per household$_{it}$)) + β2(PCA composite metric of four deprivation metrics$_{it}$) + β3(ln(total households connected$_{it}$)) + β4(% of households that only have waste connections$_{it}$) + β5(covid_2020$_{t}$) + ε_{it}

5. ln(DC_hh$_{it}$)= α + β1(ln (average bill size per household$_{it}$)) + β2(PCA composite metric of four deprivation metrics$_{it}$) + β3(ln(total households connected$_{it}$)) + β4(% of households that only have waste connections$_{it}$) + β5(covid_2020$_{t}$)+β6(covid_2021$_{t}$) + ε_{it}

6. ln(DC_hh$_{it}$)= α + β1(ln (average bill size per household$_{it}$)) + β2(simple arithmetic mean of four deprivation metrics$_{it}$) + β3(ln(total households connected$_{it}$)) + β4(% of households that only have waste connections$_{it}$) + β5(covid_2020$_{t}$) + ε_{it}

7. ln(DC_hh$_{it}$)= α + β1(ln (average bill size per household$_{it}$)) + β2(simple arithmetic mean of four deprivation metrics$_{it}$) + β3(ln(total households connected$_{it}$)) + β4(% of households that only have waste connections$_{it}$) + β5(covid_2020$_{t}$)+β6(covid_2021$_{t}$) + ε_{it}

<u>Total Cost Model</u>

8. ln(sTC_hh$_{it}$)= α + β1(ln(average bill size per household$_{it}$)) + β2(% of households with metered connections$_{it}$) + β3(ln(total households connected$_{it}$)) + β4(totalmigration$_{it}$) + β5 (PCA composite metric of four deprivation metrics$_{it}$) + β6(covid_2020$_{t}$)+β6(covid_2021$_{t}$) + ε_{it}

Description of the dependent variable

SWBRDC1, SWBRDC2, SWBRDC3, SWBRDC4: bad debt related cost per household (ratio between DC_t and hh_t, with both variables extracted from realstatafile in the published retail dataset).

SWBRTC1: total cost per household with smoothed depreciation (ratio between sTC_tr and hh_t, with both variables extracted from realsttafile in the published retail dataset).

SWBROC1: total cost less debt per household with smoothed depreciation (ratio between sOC_tr and hh_t, with both variables extracted from realstatafile in the published retail dataset).

SWBROC2 and SWBROC3: we have departed from Ofwat's dependant variables and established other costs on a *per service* basis as well. The dependent variable 'sOC_ss' is explained in the table below.

Identifier	**Description**	**Calculation notes**
sOC_ss	Other opex for modelling at a total level (code: sOC_tr in realstatafile, as published retail dataset), divided by total weighted* services connected. **the weighting is explained further below in the section covering explanatory variables*	= sOC_tr/ (R3017 + R3019 + R3018 + R3020 + R3022_weighted + R3021_weighted)

Description of the explanatory variables

We have added in new explanatory variables in both bottom-up and top-down models.

On the bad debt side, this has involved the inclusion of composite variables which account for all relevant deprivation indicators.

On the other retail costs side, we have converted certain explanatory variables to a *per service* basis which were previously in Ofwat's models on a *per household* basis.

We have also included covid dummy variables and a timetrend variable.

Explanatory Variables		
Identifier	**Description**	**Calculation notes**
covid_2020	A dummy variable for FY2019-2020	= (year == 2020)
covid_2021	A dummy variable for FY2020-2021	= (year == 2021)
timetrend	time trend starting at 1 in 2013/14 (i.e. 1 in 2013/14, 2 in 2014/15, 3 in 2015/16, etc.).	
hh_t	Total households connected (code: hh_t in real statafile), as reported in the published retail dataset.	
hhwaste_hh	% waste/sewerage only service connections, using variables R3019, R3020 and hh_t in realstatafile, as reported in the published retail dataset.	= (R3019 + R3020) / hh_t * 100
comp_pca_4a	Constructed using principal components analysis, based on all 3 Equifax variables (% of households with default, credit risk score (inverse), and court judgements per household) and the IMD Income deprivation metric (the interpolated version) as published in the retail dataset. Detailed in the next column.	See the supporting do files for all the details on how this measure has been constructed.
comp_arith_4a	Simple arithmetic mean of all 3 Equifax variables (% of households with default, credit risk score (inverse), and court judgements per household) and IMD Income deprivation metric (the interpolated version) as published in the retail dataset – after all variables have been standardised *Note: 'standardised' implies that the original variable is first de-meaned and*	See the supporting do files for all the details on how this measure has been constructed.

	normalised by dividing it through its respective standard-deviation - as to created comparable metrics that are on the same scale	

R3021_weighted	Constructed using the variable 'Households connected for water and sewerage – unmetered' (code: R3021 in statarealfile as published in the retail data set), with different weights applied for ROC2 and ROC3 to account for the different costs in serving customers with dual services	ROC2:
		R3021_weighted = R3021 * 2.0
		ROC3:
		R3021_weighted = R3021 * 1.3

R3022_weighted	Constructed using the variable 'Households connected for water and sewerage – metered' (code: R3022 in statarealfile as published in the retail data set), with different weights applied for ROC2 and ROC3 to account for the different costs in serving customers with dual services	ROC2:
		R3022_weighted = R3022 * 2.0
		ROC3:
		R3022_weighted = R3022 * 1.3

hhm_hh_s	% of metered services connected, using weighted data from published retail dataset as indicated in the next column	hhm_hh_s =(R3018 + R3020 + R3022_weighted) / hh_s * 100
hh_s_ww	% of total services that are wastewater, using weighted data from published retail dataset as indicated in the next column. Dual service households are divided in half to isolate wastewater services.	hh_s_ww = (R3019 + (R3021_weighted * 0.5) + (R3022_weighted * 0.5) + R3020) / hh_s * 100
hh_s	Total number of connected services based on weighted data as reported in the published retail dataset as indicated in the next column	hh_s =R3017 + R3019 + R3018 + R3020 + R3022_weighted + R3021_weighted
Total migration	% of total internal + international migration (code: totalmigration in statarealfile) as reported in the published retail dataset	

Brief comment on the models

General Comments

- **Time period & structure:** The modelling period is 2014-2022.
  - The only deviations from the published do-file is the inclusion of covid dummy variables for 2020 and 2021.
  - Some combination of covid dummies and time trends have been applied in all models, as and where they prove significant and operationally justified.
- **Relative performance**: All models included perform well in terms of statistical significance and on the specified tests. In particular, the suggested models all outperform the original PR19 models in terms of goodness of fit and the significance of the coefficients of the relevant explanatory variables.
  - Note: We have included an additional measure of goodness of fit, the Root Mean Squared Error (RMSE), as an additional metric. One benefit of this metric is that allows one to compare model fit across models in the same category (even if the dependent variable has been changed – as in the case of the Other operating costs presented here).
- Note our suggested solution to correlation among potential deprivation metrics in bad debt and total cost models: **composite deprivation metrics**
  - We construct and introduce composite deprivation metrics, to proxy for the probability of default. This metric thus substitutes the original variables used by Ofwat to the same end (eq_lpcf62 and the IMD Income score (unadjusted version)). These variables are thus used in both the disaggregated bad-debt models and related, top-down total cost model.
  - **The rationale** for the composite deprivation metrics is that it avoids both (i) the 'cherry picking' individual deprivation metrics and (ii) collinearity issues (if multiple suitable metrics were to be included individually).
  - **comp_pca_4a** : Principal Component Analysis (pca) was used to select the relevant variables to include into the deprivation composite metric, including variables based on their objective statistical properties (as identified in the pca). In sum, a combination of all 3 Equifax variables and the IMD Income score (interpolated) allows one to construct a variable that is both internally consistent and maximises the variance of the underlying variable. Note:
    - We use the interpolated version of the IMD income score, because it is i) slightly more highly correlated with the outcome variable of interest and (ii) has a higher item-rest correlation with the other 3 Equifax variables.
    - The council tax variable is excluded from the pca construction, due to its low item-rest correlation and high uniqueness.
  - **comp_arith_4a**: should a more simple combination (and more intuitive to explain the composite version) of the variable be preferred, we also include a similar composite metric constructed by taking the arithmetic mean of the

standardised versions of the same 4 variables (as identified by means of the broader pca analysis).
    - **The results are almost equivalent**, independent of which of the two composite metrics are preferred (and so both are included, for completeness).

Below the relevant comments for each respective bottom up and top down model:

<u>Other Retail Costs</u>

- Other retail costs have been modelled on both a 'per household' (as in Ofwat's models) and on a 'per service' basis.
- The per household model seems to perform slightly better in terms of model fit, relative to the per service models (as indicated by both the RMSE metrics and max-min ranges of the resulting efficiency scores).
- With respect to the two per service models (ROC2 and ROC3):
    - a weight of between 1.0 and 2.0 must be used to account for households with dual services – we note that it was previously estimated at 1.3.
    - ROC2 uses a 2.0 weight while ROC3 uses a 1.3 weight.
    - ROC2 performs better on the specified tests, however the max-min range of the efficiency scores are roughly similar between the two models.
- We have added total migration as a cost driver of other retail costs, due to higher cost to serve customers with higher levels of household transience. This improves the model fit and significance of all the variables in the model.
- A variable accounting for economies of scale has also been included, as the greater the number of connects (and therefore customers) the less costly it is to service these customers.

<u>Bad Debt Costs</u>

- Bad debt models use Ofwat's approach of modelling on a per household basis.
- Deprivation composite variables are presented to avoid picking one deprivation variable over another and accounting for any collinearity between the deprivation variables
    - A principal components analysis is presented (see above for rationale under general comments) as well as a simple arithmetic mean
- A variable accounting for "'share of households that only have waste connections" has been included to capture the dynamic where WOCs provide bad debt services on behalf of their partnering WASC. In such cases, there is often a joint account managed by the client-facing WOC, who may provide customer services, debt management and/or meter reading, etc., on behalf of their partnering WASC (who provides the waste service itself).

- A variable accounting for economies of scale has also been included as the greater the number of connects (and therefore customers) the less costly it is to collect additional debt.

Total Cost Model

- Ofwat's dependent variable has been retained (total cost per household)
- To align with the disaggregated models above, the following variables have been included:
    - composite deprivation measure
    - total migration,
    - the number of total households connected (scale variable),
    - the average bill size per household,
    - the % of households with metered connections, as well as
    - dummy variables for financial years 2019/20 and 2020/21.

Other Retail Costs

	SWBROC1	SWBROC2	SWBROC3
Dependent variable	lnsOC_hh	lnsOC_ss	lnsOC_ss
hhdu_hh	0.007*** {0.000}		
hhm_hh	0.009*** {0.004}		
lnhh_t	-0.158*** {0.003}		
hhm_hh_s		0.008*** {0.007}	0.008** {0.011}
hh_s_ww		-0.003 {0.274}	0.001 {0.789}
lnhh_s		-0.108** {0.013}	-0.067 {0.220}
totalmigration	0.062*** {0.000}	0.049*** {0.000}	0.038*** {0.001}
hhwaste_hh	-0.004 {0.191}		
covid_2020	0.091*** {0.000}	0.086*** {0.000}	0.081*** {0.000}
timetrend	-0.021** {0.027}	-0.019** {0.030}	-0.020** {0.023}
Constant	3.598*** {0.000}	3.209*** {0.000}	2.851*** {0.000}
Estimation method (OLS or RE)	RE	RE	RE

N (sample size)	153	153	153
R2 adjusted	0.207	0.52	0.089
RMSE	0.107	0.108	0.11
RESET test	0.287	0.547	0.117
VIF (max) – OLS	4.198	3.414	2.846
Pooling / Chow test	1.00	1.00	1.00
Normality of model residuals – OLS	0.399	0.359	0.581
Heteroskedasticity of model residuals – OLS	0.886	0.058	0.002
Test of pooled OLS versus Random Effects (LM test)	0	0	0
Efficiency score distribution (min and max)	Max: 135% Min: 76%	Max: 139% Min: 77%	Max: 140% Min: 75%
Sensitivity of estimated coefficients to removal of most and least efficient company	A	G	G
Sensitivity of estimated coefficients to removal of first and last year of the sample	A	A	A

Bad Debt Costs

	SWBRDC1	**SWBRDC2**	**SWBRDC3**	**SWBRDC4**
Dependent variable	lnDC_hh	lnDC_hh	lnDC_hh	lnDC_hh
lnrev_hh	1.268*** {0.000}	1.305*** {0.000}	1.266*** {0.000}	1.302*** {0.000}
comp_pca_4a	0.262*** {0.000}	0.296*** {0.000}		
comp_arith_4a			0.281*** {0.000}	0.318*** {0.000}
lnhh_t	-0.232*** {0.002}	-0.263*** {0.000}	-0.231*** {0.002}	-0.262*** {0.000}
hhwaste_hh	0.011*** {0.001}	0.012*** {0.000}	0.011*** {0.001}	0.012*** {0.000}
covid_2020	0.395*** {0.000}	0.438*** {0.000}	0.395*** {0.000}	0.438*** {0.000}

covid_2021		0.238*** {0.005}		0.238*** {0.005}
Constant	-1.741*** {0.002}	-1.545*** {0.003}	-1.737*** {0.002}	-1.541*** {0.003}
Estimation method (OLS or RE)	RE	RE	RE	RE
N (sample size)	153	153	153	153
R2 adjusted	0.717	0.73	0.717	0.73
RMSE	0.31	0.301	0.31	0.301
RESET test	0.065	0.066	0.065	0.066
VIF (max) - OLS	2.949	2.971	2.938	2.96
Pooling / Chow test	0.934	0.999	0.933	0.999
Normality of model residuals - OLS	0	0	0	0
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 and max)	Max: 160% Min: 82%	Max: 161% Min: 80%	Max: 160% Min: 82%	Max: 161% Min: 80%
Sensitivity of estimated coefficients to removal of most and least efficient company	G	A	G	A
Sensitivity of estimated coefficients to removal of first and last year of the sample	A	G	A	G

Total Costs

	SWBRTC1
Dependent variable	lnsTC_hh
lnrev_hh	0.745*** {0.000}
hhm_hh	0.005 {0.157}
lnhh_t	-0.180*** {0.000}
totalmigration	0.047*** {0.002}
comp_pca_4a	0.150*** {0.001}
covid_2020	0.201*** {0.000}

covid_2021	0.064** {0.020}
Constant	0.824*** {0.005}
Estimation method (OLS or RE)	RE
N (sample size)	153
R2 adjusted	0.718
RMSE	0.109
RESET test	0.327
VIF (max) - OLS	5.24
Pooling / Chow test - OLS	1
Normality of model residuals - OLS	0.146
Heteroskedasticity of model residuals - OLS	0.024
Test of pooled OLS versus Random Effects (LM test)	0
Efficiency score distribution (min and max)	Max: 123% Min: 81%
Sensitivity of estimated coefficients to removal of most and least efficient company	G
Sensitivity of estimated coefficients to removal of first and last year of the sample	A

Efficiency scores SWBROC1

Company	Rank	Efficiency score
SWB	1	76%
BRL	2	78%
ANH	3	79%
NWT	4	89%
SEW	5	89%
WSX	6	92%
AFW	7	92%
PRT	8	93%
YKY	9	95%
HDD	10	106%
TMS	11	106%
SVE	12	114%
SSC	13	116%
SRN	14	116%
WSH	15	121%

NES	16	130%
SES	17	135%

Efficiency scores SWBROC2

Company	Rank	Efficiency score
ANH	1	77%
BRL	2	79%
SWB	3	83%
WSX	4	84%
SEW	5	88%
AFW	6	89%
PRT	7	92%
NWT	8	97%
YKY	9	102%
TMS	10	106%
SSC	11	108%
HDD	12	108%
SRN	13	109%
SVE	14	112%
WSH	15	128%
NES	16	130%
SES	17	139%

Efficiency scores SWBROC3

	Rank	ROC3
ANH	1	75%
WSX	2	77%
BRL	3	80%
SEW	4	85%
AFW	5	86%
SWB	6	91%
PRT	7	92%
SSC	8	102%
SRN	9	103%
NWT	10	104%
TMS	11	104%
YKY	12	109%
HDD	13	111%
SVE	14	112%
NES	15	129%
WSH	16	135%
SES	17	140%

Efficiency scores SWBRDC1

Company	Rank	Efficiency score
SVE	1	82%

SWB	2	86%
NES	3	87%
YKY	4	89%
SEW	5	92%
WSX	6	97%
SES	7	99%
PRT	8	99%
NWT	9	101%
ANH	10	103%
WSH	11	108%
TMS	12	115%
HDD	13	121%
SRN	14	123%
SSC	15	141%
AFW	16	149%
BRL	17	160%

Efficiency scores SWBRDC2

Company	Rank	Efficiency score
SVE	1	80%
NES	2	83%
SWB	3	85%
YKY	4	86%
NWT	5	95%
WSX	6	96%
SEW	7	96%
PRT	8	98%
SES	9	99%
ANH	10	102%
WSH	11	103%
HDD	12	112%
TMS	13	113%
SRN	14	120%
SSC	15	137%
AFW	16	151%
BRL	17	161%

Efficiency scores SWBRDC3

Company	Rank	Efficiency score
SVE	1	82%
SWB	2	85%
NES	3	86%
YKY	4	89%
SEW	5	92%
WSX	6	97%
SES	7	99%
PRT	8	99%

NWT	9	101%
ANH	10	103%
WSH	11	108%
TMS	12	115%
HDD	13	121%
SRN	14	123%
SSC	15	141%
AFW	16	149%
BRL	17	160%

Efficiency scores SWBRDC4

Company	Rank	Efficiency score
SVE	1	80%
NES	2	82%
SWB	3	85%
YKY	4	86%
NWT	5	95%
WSX	6	96%
SEW	7	96%
PRT	8	98%
SES	9	99%
ANH	10	102%
WSH	11	103%
HDD	12	112%
TMS	13	113%
SRN	14	120%
SSC	15	137%
AFW	16	151%
BRL	17	161%

Efficiency scores SWBRTC1

Company	Rank	Efficiency score
SWB	1	81%
BRL	2	88%
SEW	3	89%
ANH	4	92%
AFW	5	92%
NWT	6	93%
YKY	7	95%
NES	8	100%
PRT	9	100%
WSX	10	101%

SVE	11	101%
TMS	12	104%
SSC	13	108%
HDD	14	110%
WSH	15	118%
SES	16	120%
SRN	17	123%