Reckon LLP

Data analysis | Economic regulation | Competition law

Capturing deprivation and arrears risk in household retail cost assessment

Working paper for United Utilities on Wednesday 10 May 2017

Table of contents

Reckon LLP, limited liability partnership registered in England (number OC307897)
31 Southampton Row, London, WC1B 5HJ • www.reckon.co.uk

Introduction and summary

Context

1. Previous work in the water industry has treated economic and social deprivation as one of the drivers of companies' bad debt costs. In particular, in its PR14 final determinations, Ofwat reflected, albeit with some modifications, proposals from several companies for an upward financial adjustment to the household retail cost to serve allowance (derived from industry-wide retail cost benchmarking) to take account of greater levels of deprivation in their areas of appointment relative to other companies. For example, in its assessment of United Utilities' proposal, Ofwat concluded that "United Utilities provided sufficient and convincing evidence that deprivation (especially extreme deprivation as measured by the 10 per cent most deprived households) affects United Utilities in a materially different way to other companies".[1] The evidence base for United Utilities' proposal drew on data available from the ONS relating to the index of multiple deprivation (IMD).

2. The ONS deprivation data provides a rich characterisation of deprivation across England at a geographically granular level. The use of this data for cost assessment models is, however, constrained by two shortcomings. First, the ONS data only covers England and does not cover Wales. (Statistics for Wales publishes similar deprivation measures, though these are not entirely consistent with those produced by the ONS). Second, the ONS data is published only every few years; the last three versions were published in 2015, 2010 and 2007.

3. We are not aware of alternative published data that captures deprivation at the local level, and from which measures at the level of the areas served by water companies in England and Wales could be constructed.

4. For PR19, there is an opportunity to address this. United Unities has been working with Equifax to identify additional sources of relevant data, which could help tackle some of the limitations of the deprivation data available from the ONS.

5. Reckon has been supporting United Utilities with analysis of the data provided by Equifax. We have sought to identify good quality candidate variables to reflect deprivation and arrears risk in the water sector. These variables can be used by Ofwat and companies for the purpose of cost assessment and for explaining differences in debt cost across companies. The work has been structured into two phases. We summarise below the work we have done to date and draw out some emerging findings.

Summary of Phase One to derive shortlist of Equifax variables

6. In Phase One of the work programme, United Utilities commissioned Reckon to carry out quantitative analysis on a sample dataset provided by Equifax. The sample of Equifax data that we used covered anonymised postcodes in 269 Lower Layer Super Output Areas (LSOAs) in England. These LSOAs represented around one per cent of the total number of LSOAs in England. The Equifax sample dataset contained a very

[1] Ofwat (2014) *Draft price control determination notice: company-specific appendix – United Utilities*, page 33

large number of variables relating to the characteristics, credit history and credit risk of households in the postcodes.

7. The key steps we took in Phase One were as follows:

   (a) Following receipt of the data on the Equifax variables, we carried out an initial stock-take of the available data on these variables and made proposals to United Utilities as to which should be included in the quantitative analysis. We agreed on the exclusion of variables that did not seem meaningful or useful for the purposes of the project (for example, variables relating to age distribution or to marital status). This process resulted in a set of around 400 Equifax variables being taken forward to our quantitative analysis.

   (b) We carried out some data processing to convert the raw data from Equifax at the postcode level to data that could be compared with (i) measures of deprivation available from the ONS for LSOAs in England and (ii) approximate allocations of United Utilities' bad debt costs and other household retail costs between LSOAs in the geographic area it supplies.

   (c) The main part of our analysis involved taking each Equifax variable in turn and running a series of econometric regressions involving that variable as an explanatory variable in the regression model. We used a variety of dependent variables, covering both ONS deprivation measures and measures of United Utilities' bad debt costs at the LSOA level. The regression results provided measures of goodness of fit between the Equifax explanatory variables and the dependent variables and hence of the degree of correlation between them.

8. The outcome of Phase One was our suggested shortlist of Equifax variables that seemed particularly promising for subsequent analysis. The list primarily reflected the estimation results from the econometric regressions, as well as our thoughts on the intuitive rationale for the Equifax variables, and a desire to identify a set of variables that captured a good range of different factors. A degree of judgement was involved. United Utilities contributed to the overall selection, drawing on a review of the regression results and on its operational insight.

Main strands of work for Phase Two

9. Following Phase One, United Utilities procured a dataset from Equifax containing information on the 29 shortlisted Equifax variables across all postcodes in England and Wales, spanning the period 2006 to 2015. United Utilities asked us to explore the use of this dataset for the purpose of understanding the drivers of water company bad debt and other household retail costs, and developing econometric models covering water companies in England and Wales.

10. At the start of Phase Two, we confirmed that the results from Phase One still held when the analysis was done on the expanded Equifax dataset which covered all LSOAs in England and in Wales. We found that the high correlations that we had found in Phase One between the shortlisted Equifax variables and measures of deprivation from Phase

One held when we used the expanded Equifax dataset.[2] We also confirmed that these high correlations applied when we looked separately at each of England, Wales and London and at the different years covered by the dataset.

11. The remainder of our work during Phase Two involved the following:

(a) We used the expanded Equifax dataset to calculate the weighted-average value of each Equifax variable for each of the water companies in England and Wales. This provides insight on differences between companies in terms of measures of deprivation and arrears risk in the geographic areas that they supply.

(b) We used the Equifax dataset to develop econometric models that would enable us to construct proxy or predicted values of the ONS measures of deprivation for LSOAs across England and Wales and across different years. This allowed us to remedy coverage issues due to the incompatibility of the ONS measures of deprivation in England and the analogous measures produced by Statistics for Wales.

(c) United Utilities provided us with a dataset which gave an estimated breakdown of its retail costs of serving households across the 4,500 or so LSOAs that make up the region it covers. We combined that information with the Equifax dataset and with data on ONS measures of deprivation to develop econometric models that sought to explore the extent to which variations in United Utilities' debt costs across the LSOAs could be explained by differences with respect to ONS and Equifax variables.

(d) We drew on the strands of work above to develop and test econometric models of company-level water debt costs for 18 water companies in England and Wales.[3] In developing those models, we sought to examine the degree to which variations in the levels of deprivation and arrears risk in the areas served by companies could explain differences in their debt costs.

Findings

12. Phase Two provides our first application of the data from Equifax to water company retail cost benchmarking. We summarise below our findings from the work so far.

13. We were able to develop econometric models to construct proxy or predicted values of ONS measures of deprivation using Equifax data. The explanatory variables in the models included variables relating to socio-economic and demographic characteristics of the population in each LSOA (e.g. employment status and qualifications) and variables relevant to the arrears risk of households in the LSOA (e.g. an Equifax proprietary risk score and a variable measuring prevalence of County Court Judgments for debt). Table 1 lists the six Equifax variables included in our preferred model based on work to date.

[2] A minority of the variables selected from Phase One had showed relatively low correlations in that initial phase, and we found similar results for these in Phase Two.
[3] For the purpose of our analysis we treated Bournemouth Water and South West Water as separate.

Table 1 Equifax variables used to construct proxy ONS measures of deprivation

Reference	Variable description
LPCF72	Percentage of households with zero reported payment issues in the last six months
RGC102	Equifax proprietary credit risk score
XPCF2	Average number of County Court Judgments per household
GCG543	Percentage of population with no educational qualifications
GCG557	Percentage of population that are inactive for employment purposes due to sickness
MGC191	Percentage of households in Council Tax Band A

14. The models developed using these variables provided a good fit to the data. The R-squared statistics were above 0.9 indicating that over 90 per cent of the variation in the ONS measure of deprivation across the LSOAs are explained by the variation in the value of the Equifax variables used in the models. We applied these models to obtain predicted values of the ONS measures for all LSOAs in England and Wales, in each year, and, from that, to aggregate these "predicted ONS" measures to the water company level.

15. The econometric models we developed of United Utilities' water debt cost across LSOAs suggest that differences in deprivation measures do explain observed variation in costs. We found that models that use Equifax variables to control for deprivation can provide a better fit of the data than ones that draw only on ONS measures of deprivation. We found the R-squared statistics of models that included only ONS deprivation models to be around 0.60, whilst for a models that included Equifax as alternative explanatory variables this statistic was around 0.73.

16. From our company-level modelling of water debt costs, we found that variables relating to deprivation, arrears risk and average bill size helped explain variations of bad debt costs across water companies. We have produced example models that use either our predicted ONS deprivation measures (derived from econometric modelling using the Equifax variables) or the Equifax variables directly. These models seem to give intuitively reasonable results. Furthermore, we applied to the models a series of diagnostic tests, ones that PwC had regard to in its review of the econometric models of doubtful debt which companies put forward at PR14. The tests detected no concern in most of the models presented.

17. The example models we have explored so far indicate it should be possible to estimate each water company's (efficient level of) bad debt costs, based on historical data across the industry and taking account of factors affecting bad debt costs such as average bill size, deprivation and the arrears risk. We summarise in Table 2 our suggestions, based on work to date, for the specification of the dependent variable and explanatory variables in company-level models.

Table 2 Suggestions on specification for company-level models emerging from Phase Two

Dependant variable	Explanatory variables
• Bad debt costs per unique customer (natural logarithm)	• Measure of average household bill • Measures of average deprivation levels across LSOAs served by each water company. We suggest constructing these measures on basis of (i) the predicted ONS IMD or (ii) the predicted ONS Income Deprivation score, both derived from econometric models using Equifax data. • Measure of average arrears risk across LSOAs served by company calculated on basis of Equifax proprietary measure of credit risk (RGC102). • Measures of incidence of "extreme" deprivation across LSOAs served by each water company. We suggest constructing these measures as the proportion of households in a company's region which are in the 10% or 20% most deprived LSOAs across England and Wales, as measured by the predicted ONS IMD, the predicted ONS Income Deprivation score or by an Equifax proprietary measure of credit risk.
• Ratio of bad debt costs to household revenues	• As above, but excluding the measure of average household bill

18. We explored different specifications and approaches for the company-level models. Deprivation/arrears risk correlates with both debt-related costs as well as all retail costs. However, the t-statistics on the estimated coefficients for the deprivation/arrears risk variables tended to be lower in the models covering all retail operating costs than in models focused on debt costs, indicating more imprecision in the former.

19. Overall, we believe that Phase Two demonstrates grounds for using variables derived from the Equifax dataset as part of household retail cost assessment, and shows how these variables can be successfully incorporated into econometric benchmarking models.

20. This paper sets out the progress of the work carried out so far. It is not intended to present a final set of preferred models or variables. We expect that further work could bring additional insight and benefit, for example by refining the way that variables derived from the Equifax dataset are used in the model specifications for company-level econometric benchmarking models.

Structure of the paper

21. The remainder of this paper is structured as follows:

    (a) We introduce the ONS and the Statistics for Wales measures of deprivation that we have considered and compare the weighted averages of the ONS measures across water companies in England.

(b) We introduce the Equifax variables that are available from the Phase Two dataset and compare the weighted averages of these across water companies in England and Wales.

(c) We present econometric analysis of how variations in the ONS deprivation measures and the Equifax variables at LSOA-level can explain variation in the levels of United Utilities' bad debt costs across the LSOAs within its area of appointment.

(d) We present analysis of the association between the Equifax variables and ONS measures of deprivation, including econometric modelling to predict ONS deprivation measures from the Equifax dataset.

(e) We present econometric analysis comparing measures of bad debt across water companies, drawing on the Equifax dataset.

The ONS and Statistics for Wales deprivation measures

22. Several of the strands of analysis we are concerned with draw on measures of deprivation published by the ONS, for England, and by Statistics for Wales.

23. In particular, the two agencies construct and publish on a regular basis (every 3 to 5 years) an Index of Multiple Deprivation (IMD). The IMD is constructed at the level of the Lower Layer Super Output Areas (LSOAs) and, in broad terms, is a weighted average of the ranking of the LSOAs in each of several domains of deprivation. Of the domains taken into account, "income deprivation" and "employment deprivation" are the two that are given the most weight. In the case of the ONS calculation of the IMD for 2015, each of those domains was given a weight of 22.5 percent. Other domains of deprivation that are taken into account are education, health, crime, barriers to housing and services and living environment deprivation.

24. Of the domains of deprivation considered by ONS/Statistics for Wales, those that appear most relevant in the context of analysing the association of deprivation and the debt costs of water companies are income deprivation, employment deprivation and the IMD itself. We focus on these.

25. In the rest of this section, we first outline how ONS/Statistics for Wales define each of these three measures and, following from that, comment on the non-comparability of the measures between Wales and England, and across time. We then compare the water companies in England in terms of the three deprivation measures of the LSOAs within the area they serve.

Definition of deprivation scores

26. The measures of income and of employment deprivation are based on the percentage of the population, within each LSOA, that meet one or more of a set of criteria. The criteria used by the ONS to construct these measures for England are outlined in Table 3.[4]

Table 3 Deriving the income and employment deprivation scores 2015 (ONS)

Income deprivation domain

Measures proportion of the population in an area experiencing deprivation relating to low income. The definition of low income used includes both those people that are out-of-work, and those that are in work but who have low earnings (and who satisfy the respective means test).

The measure is calculated as proportion of population who satisfy one or more of the following:

- Adult and children in Income Support families
- Adults and children in income-based Jobseeker's Allowance families
- Adults and children in income-based Employment and Support Allowance families
- Adults and children in Pension Credit (Guarantee) families
- Adults and children in Working Tax Credit and Child Tax Credit families not already counted, that is those who are not in receipt of Income Support, income-based Jobseeker's Allowance, income-based

[4] ONS (2015) "The English Indices of Deprivation 2015, Technical report".

Employment and Support Allowance or Pension Credit (Guarantee) and whose equivalised income (excluding housing benefit) is below 60 per cent of the median before housing costs • Asylum seekers in England in receipt of subsistence support, accommodation support, or both
Employment deprivation domain Measures proportion of the working-age population involuntarily excluded from the labour market. Includes those who would like to work but are unable to do so due to unemployment, sickness or disability, or caring responsibilities. The measure is calculated as the proportion of working-age population who satisfy one or more of the following: • Claimants of Jobseeker's Allowance (both contribution-based and income-based), women aged 18-59 and men aged 18-64 • Claimants of Employment and Support Allowance (both contribution-based and income-based), women aged 18-59 and men aged 18-64 • Claimants of Incapacity Benefit, women aged 18-59 and men aged 18-64 • Claimants of Severe Disablement Allowance, women aged 18-59 and men aged 18-64 • Claimants of Carer's Allowance, women aged 18-59 and men aged 18-64

27. In broad terms, the ONS derives the Index of Multiple Deprivation (IMD) of an LSOA as a weighted average of the ranking of the LSOA across the various domains of deprivation mentioned earlier. To appreciate the IMD, and to inform on how analyses that draw on it can be interpreted, it is useful to outline the main steps involved in deriving that measure:[5]

    (a) For each domain of deprivation, the ONS constructs a score for each LSOA. In the case of the income deprivation domain and the employment deprivation domain, the score is the percentage of households who meet at least one of a number of conditions (relating to income, or to employment), as outlined in Table 3. In the case of the other domains, the score involves bringing together measures across a number of indicators.

    (b) For each domain, ONS ranks the LSOAs on the basis of that score. The LSOA with the lowest score is ranked 1 (for the least deprived), and the LSOA with the highest score is ranked 32,844 (the number of LSOAs in England, as of when the ONS compiled the 2015 deprivation measures).

    (c) For each domain, the ranking of LSOAs is standardised and transformed so that they have a number of features which ONS considers are appropriate for the purpose of subsequently combining the transformed ranks across domains. The standardisation and transformation is such that, for each domain, the least deprived LSOA is attributed a transformed ranking of 1, and the most deprived a transformed ranking of 100. The transformation "stretches" the range spanned by the more deprived LSOAs. For example, the transformation means that, were an LSOA to rank as the LSOA on the "border" of the top decile — i.e. the LSOA such

[5] ONS (2015) "The English Indices of Deprivation 2015, Technical report". The description of the transforming the ranks of each domain is set out in Appendix F.

that 90 per cent of LSOA are less deprived than it and 10 per cent are more deprived than it — then its transformed ranking would be 50.

(d) ONS calculates the IMD as the weighted average of the transformed ranks across domains, using weights intended to capture the relative contribution that deprivation in a given domain makes to "overall deprivation". For example, the ONS gives the income deprivation and the employment deprivation domains a weight of 22.5 per cent.

28. The data used by the ONS to construct the scores for the income and the employment domain are drawn mainly from the 2012/13 financial year. The population figures used as the denominator in the calculation — for the purpose of expressing the indicators as percentages of relevant population — refer to mid-2012 population figures.

29. Statistics for Wales constructs the income and employment deprivation score and the IMD for the LSOAs in Wales along similar lines to that used by the ONS for England. The set of indicators are not, however, comparable across the nations:

(a) Because each of the IMDs published by the ONS and by Statistics for Wales are, in essence, a sort of ranking of LSOAs within England and within Wales respectively, the two indicators cannot be brought together. Put simply, knowing that a given Welsh LSOA has a rank of, say, 23 amongst all Welsh LSOAs does not allow us to know where it would fall if ranked against the LSOAs in England.

(b) The income deprivation scores produced by ONS and Statistics for Wales are not comparable, even though they both refer to the percentage of population meeting very similar criteria. In particular, for both England and Wales, the measure includes the count of families with equivalised income that is below 60 per cent of the median income in England and Wales respectively, and median income is higher in England than it is in Wales. With regard to the employment derivation score, it is possible that the Welsh and the English scores are measures of the same thing, and therefore comparable. But we are not certain that this is so. In particular, whilst the ONS measure takes account of claimants of the Carers Allowance, as outlined in the previous table, it is possible that the Statistics for Wales measure does not.[6]

30. The three ONS measures of deprivation reported are very highly correlated between themselves. This is shown in Table 4 overleaf. The high correlation between measures is not surprising: we expect unemployment to be associated with lower income. In turn, as outlined earlier, the IMD score is constructed as a weighted average of the ranking in a number of separate deprivation domains, including income and employment deprivation. And the weights on each of these two domains is 22.5 per cent, by far the more influential domains in the IMD. Similar pairwise correlations are observed for the analogous measures produced by Statistics for Wales.

[6] See Statistics for Wales (2014) "Welsh Index of Multiple Deprivation 2014 (WIMD 2014) Technical report".

Table 4 Pairwise correlation of ONS measures of deprivation

	IMD score	Income deprivation score	Employment deprivation score
IMD score	1.00		
Income deprivation score	0.97	1.00	
Employment deprivation score	0.95	0.95	1.00

Variation in deprivation scores across water companies

31. We calculated the average of the IMD and of the income and employment deprivation score published by the ONS in 2015 across the LSOAs falling within the area served by each water company in England. We calculated these as a weighted average of the scores in the relevant set of LSOA, using population in each LSOA as weights. Table 5 reports our estimate of these measures.

Table 5 Company-wide average ONS deprivation measures (based on ONS 2015 data)

Company	IMD Score	Income deprivation Score	Employment deprivation Score
AFW	16.2	12%	9%
ANH	19.0	13%	11%
BRL	18.7	12%	11%
DVW	—	—	—
NES	24.5	17%	15%
NWT	26.9	17%	15%
PRT	19.4	12%	10%
SBW	15.9	11%	9%
SES	11.7	8%	7%
SEW	12.5	9%	8%
SRN	17.8	12%	10%
SSC	21.7	15%	12%
SVT	23.3	16%	13%
SWT	21.6	13%	12%
TMS	19.3	14%	10%
WSH	—	—	—
WSX	17.2	11%	10%
YKY	25.2	16%	14%

32. The estimates in Table 5 are based on the measures published by the ONS for the English LSOAs; they do not take account of the scores published by Statistics for Wales. As such they do not take account of the deprivation in the LSOAs within Wales. As the bulk of the LSOAs served by Welsh Water and by Dee Valley Water are in Wales, and so not covered by the ONS data, we do not report figures for those two companies. Of the remaining water companies, the exclusion of Welsh LSOAs from the averages reported in the table also affects the estimate for Severn Trent, as a sizeable portion of the LSOAs that it serves are in Wales.

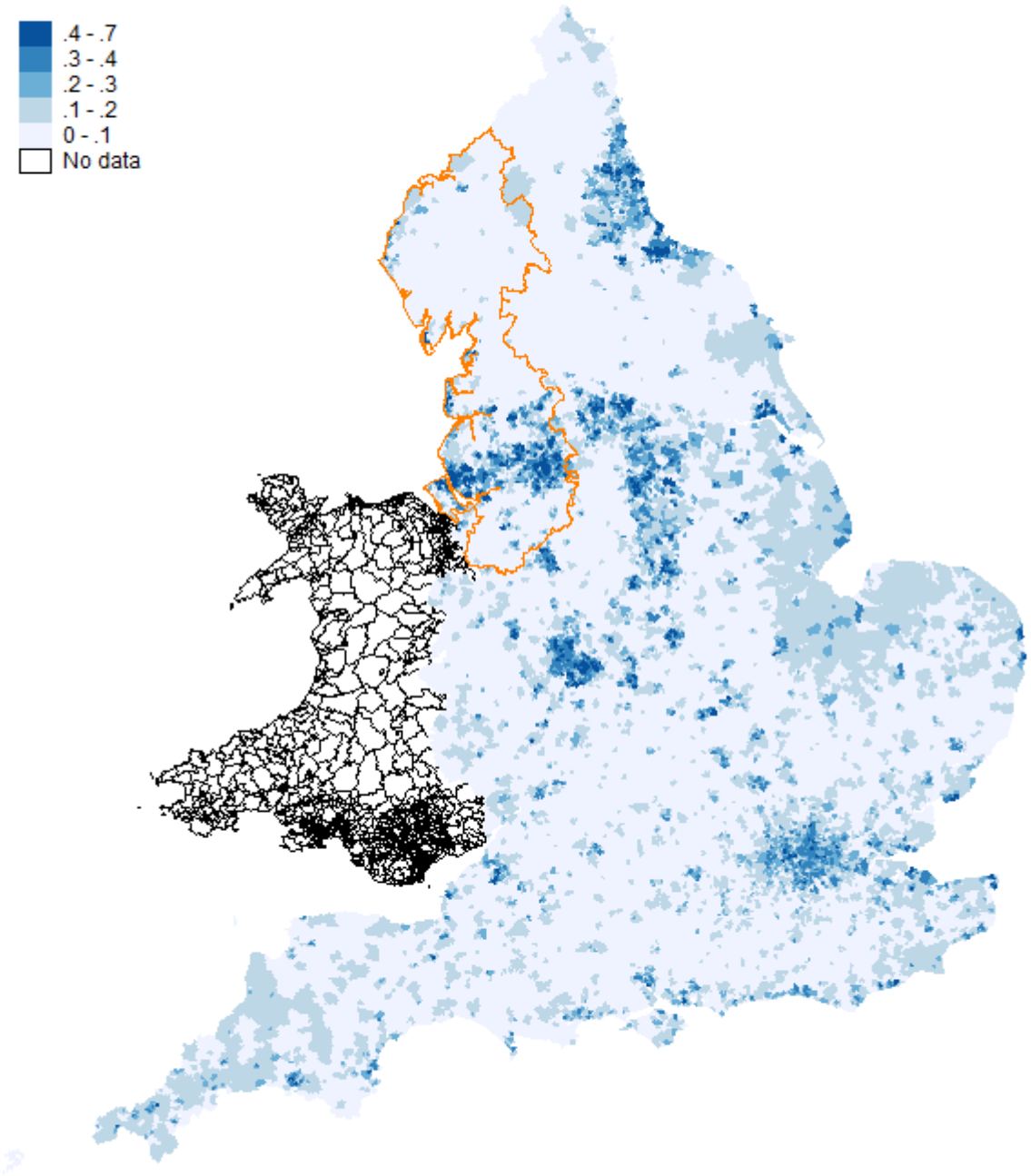
33. We should add that the figures reported in Table 5 draw on our mapping of the correspondence between LSOAs and the areas served by water companies. To carry out this mapping we have relied on two separate datasets. To map LSOAs to companies' water supply areas we used data provided by the Drinking Water Inspectorate (DWI). To map LSOAs to sewerage service areas we used a dataset circulated within water companies and Ofwat in 2016 which reports the correspondence between LSOAs and sewerage service areas.[7] Whilst the data from the DWI allowed us to make an accurate mapping between LSOAs and water supply areas, the mapping done in the dataset relating to sewerage services is done on the basis of mapping Local Authority Districts (LADs) to companies. Mapping at the LAD level does not provide as fine a level of granularity as would be desirable given that more than one wastewater company may operate within the same LAD, each serving different sets of LSOAs. We return to the discussion relating to the mapping of LSOAs to companies later on in this report.

34. Figure 1 maps the ONS income deprivation score across the LSOAs in England, providing a richer picture of the variation across England. In the map, we have highlighted the border of the region served by United Utilities. The map shows no data for Welsh LSOA, reflecting the fact that the ONS measure is not calculated for Wales.

35. The areas shaded in the darkest blue are those LSOA whose income deprivation score is above 0.4. LSOAs in this range are amongst the 2.5 per cent most deprived LSOAs according to that measure. Across the English LSOAs, the median value of the income deprivation score was 0.11.

[7] On our request, DWI kindly provided us with a dataset containing eastings and northings of water company boundaries (received on 25 April 2017). We combined this with data from the ONS to identify within which water supply area each of the LSOAs fell. To make a correspondence between LSOAs and sewerage service areas we drew on data in Excel file circulated within water companies and Ofwat (16 May 2016) in the context of work of the "Totex sub group" for PR19.

Figure 1 Map of ONS income deprivation score across English LSOAs



The Equifax dataset

36. This section gives an overview of the Equifax dataset. It presents the set of variables within it, and then outlines the variation of these variables across companies.

Overview of the Equifax dataset

37. United Utilities provided us with a dataset containing data on 29 variables compiled by Equifax. We refer to this as the Equifax dataset. The Equifax dataset reports data at the postcode level and covers the UK. The data is reported for the period 2006 to 2015, on an annual basis.

38. The 29 variables in the Equifax dataset are a subset of around 450 variables initially compiled by Equifax for United Utilities. The selection of the subset of 29 variables from that wider set, reflect the findings from an earlier phase of our work for United Utilities, where we analysed correlations between the ONS measures of deprivation and the variables in that wider set, as well as judgement on which of the variables were intuitively more reasonable in explaining variations in deprivation and/or costs associated with water debt.

39. The 29 variables cover a range of characteristics. In broad terms, and given the context of our analysis, it is useful to categorise the variables into two groups. There is one group of variables that refers to underlying socio-economic and demographic characteristics of the local area. The second group of variables relates to different measures of arrears or arrears risk compiled and/or developed by Equifax.

40. Of the 29 variables, eight relate to the proportion of the households in each of eight different Council Tax bands. Of these, for the purpose of the analysis presented below, we focused on the one relating to the percentage of households in Council Tax Band A, the lowest of the bands. This reduced to 22 the number of Equifax variables we explore in our analysis.

41. We list these 22 variables in Table 6 overleaf. We have used colour coding to indicate within which of the two groups we consider each of the variable falls in. For each variable, the table shows the time period for which data is available, and the frequency with which values are updated over that period.

Table 6 Overview of Equifax variables

Colour coding	
	Socio-economic and demographic characteristic
	Measure of arrears risk

	Variable	Available years	Frequency of update
	AGC300 – Wealth Indicator - semi-decile ranking of Wealth of Postcode (1 = High Wealth, 20 = Low Wealth)	2006 – 2015	Updated 3 times (2007, 2008 and 2014)
	AGC301 – Consumer Activity Indicator - semi-decile ranking of Consumer Activity of Postcode (1 = High Activity, 20 = Low Activity)	2006 – 2015	Updated 3 times (2007, 2008 and 2014)
	CPCF16 – CCJ Postcode Event – % Households with CCJs	2006 – 2015	Updated annually
	EPCF27 – Electoral Roll Postcode Event - average number of occupancy changes per household	2012 – 2015	Updated annually
	GCG543 – CENSUS Population Qualifications None	2006 – 2015	Updated once (2014)
	GCG552 – GCENSUS Population Employment Unemployed	2006 – 2015	Updated once (2014)
	GCG557 – CENSUS Population Employment Inactive Sick	2006 – 2015	Updated once (2014)
	GCG609 – CENSUS Household Dependant Kids and Employment Dependent Children in Household and 0 Adults in Employment	2006 – 2015	Updated once (2014)
	GCG689 – CENSUS Household Car Usage 0	2006 – 2015	Updated once (2014)
	GCG698 – CENSUS Household Tenure - Rented LA	2006 – 2015	Updated once (2014)
	LPCF18 – Full Insight Postcode Event - % households with 1 or more Credit/Store Card accounts	2006 – 2015	Updated annually
	LPCF57 – Full Insight Postcode Event - % households with total credit limit on active revolving Insight > £10,000	2006 – 2015	Updated annually
	LPCF62 – Insight Postcode Event - % households with default	2006 – 2015	Updated annually
	LPCF72 – Insight Postcode Event - % households with worst status in last 6 months active revolving Insight = 0	2006 – 2015	Updated annually
	MGC140 – Landscape Risk – Non Insight Credit Risk propensity score	2009 – 2015	Single value across years
	MGC191 – Landscape Property CT A – % of households in postcode that are Council Tax Band A	2009 – 2015	Single value across years
	RGC100 – Postcode Risk Navigator Base - Credit Risk score derived from non-Insight data	2006 – 2015	Updated annually
	RGC102 – Postcode Risk Navigator Full - Credit Risk score derived from all Insight data	2006 – 2015	Updated annually
	WGC012 – Insight Postcode Event - % households with a Credit Card account current status 3+	2009 – 2015	Single value across years
	WGC200 – Insight Postcode Event - % households with a Mail Order account current status 'D'	2009 – 2015	Updated annually
	XPCF2 – Partial Insight Postcode Event – Average number of Partial Insight accounts or CCJs per household	2006 – 2015	Updated annually
	XPCF9 – Default Insight Postcode Event - % households with 0 default accounts	2006 – 2015	Updated annually

The frequency of updates to the Equifax variables

42. We indicated in Table 6 the frequency with which the data for each of the variables in the Equifax dataset is updated in the period 2006 to 2015. Whilst for some variable the dataset reports different values for each year, there are several variables for which the values reported in the dataset are the same for the years 2006 to 2013, and the same for 2014 and 2015. This suggests that, for those variables, over that ten-year period, the data was updated once.

43. For some of these variables that were updated once over the ten-year period, the size of the adjustment is significant. For example, across the area served by United Utilities, the value of the variable defined as "GCG609 – Census Household Dependant Kids and Employment Dependent Children in Household and 0 Adults in Employment" drops from 19.4 per cent in the period 2006 to 2013, to 4.7 percent in 2014 and 2015. We have confirmed that the jump in the series is at the postcode level — the disaggregated level at which we received the data — and is not the result of the aggregation of the data that we did for our analysis.

44. The set of variables whose values change once over the ten-year period are ones that appear to be based on data collected from the census. One possible explanation for the single revision in the values reported for those variables is that the 2006 to 2013 figures are based on the 2001 census, whilst the figures for 2014 and 2015 are based on data from the 2011 census. The data from the 2011 census started to be released from mid-2012 (estimated headline population), and disaggregated, local-level, data was released from 2013 onwards.[8]

45. The frequency and timing with which these variables are revised in the Equifax dataset has some implications on our handling and interpretation of the data relating to those variables, and on our approach to the econometric modelling completed. We discuss this in the sections where we set out the econometric modelling we carried out.

Variation in levels of Equifax variables across water companies

46. The data provided by Equifax is at the postcode level. We mapped the data first to LSOAs and then to the approximate area served by each of the 18 water companies in England and Wales. For each company, we constructed weighted averages of each of the Equifax variables, using population or household numbers as weights (as was appropriate for each of the Equifax variables).

47. Figure 2 overleaf shows the spread of each of the 22 variables in the Equifax dataset across the 18 water companies. It is based on data for 2015, the most recent year in the dataset. The "whiskers" in the chart show the range between the minimum and the maximum value of each of the variables. The orange dot marks the value for United Utilities.

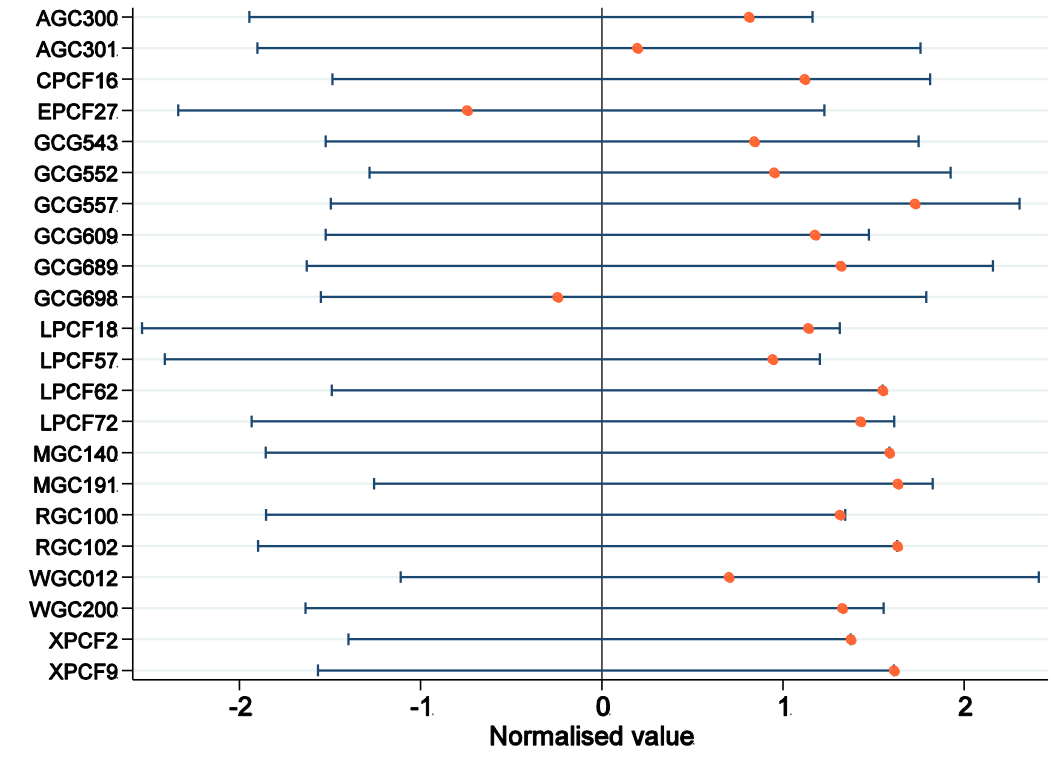
48. To bring together in the same chart the comparison for all 22 variables, we normalised their values using a standard technique. In particular, taking each variable in turn, we

[8] See http://webarchive.nationalarchives.gov.uk/20160105160709/http://www.ons.gov.uk/ons/guide-method/census/2011/census-data/2011-census-data-catalogue/census-data-quick-view/index.html

subtracted from each company's value the mean across the 18 companies and divided by the standard deviation. A normalised value of zero marks the mean; this is shown in the figure by the vertical line. A normalised value of 1, say, indicates that a company's value is one standard deviation above the industry mean. For each variable, the overall length of the whisker gives an indication of the variation in the values of that variable across the 18 companies.

49. Of the 22 Equifax variables, 16 have, plausibly, a positive association with deprivation. The remaining six are constructed or defined in such a way that, we suggest, their association with deprivation is more likely to be negative. We think this is the case of variables LPCF18, LPCF57, LPCF72, MGC140, RGC100 and RGC102.[9] Given this, and to make the reading of the chart easier to interpret, we multiplied the values for these six variables by minus one, ahead of constructing Figure 2.

Figure 2 Normalised values of Equifax variables across water companies (2015)

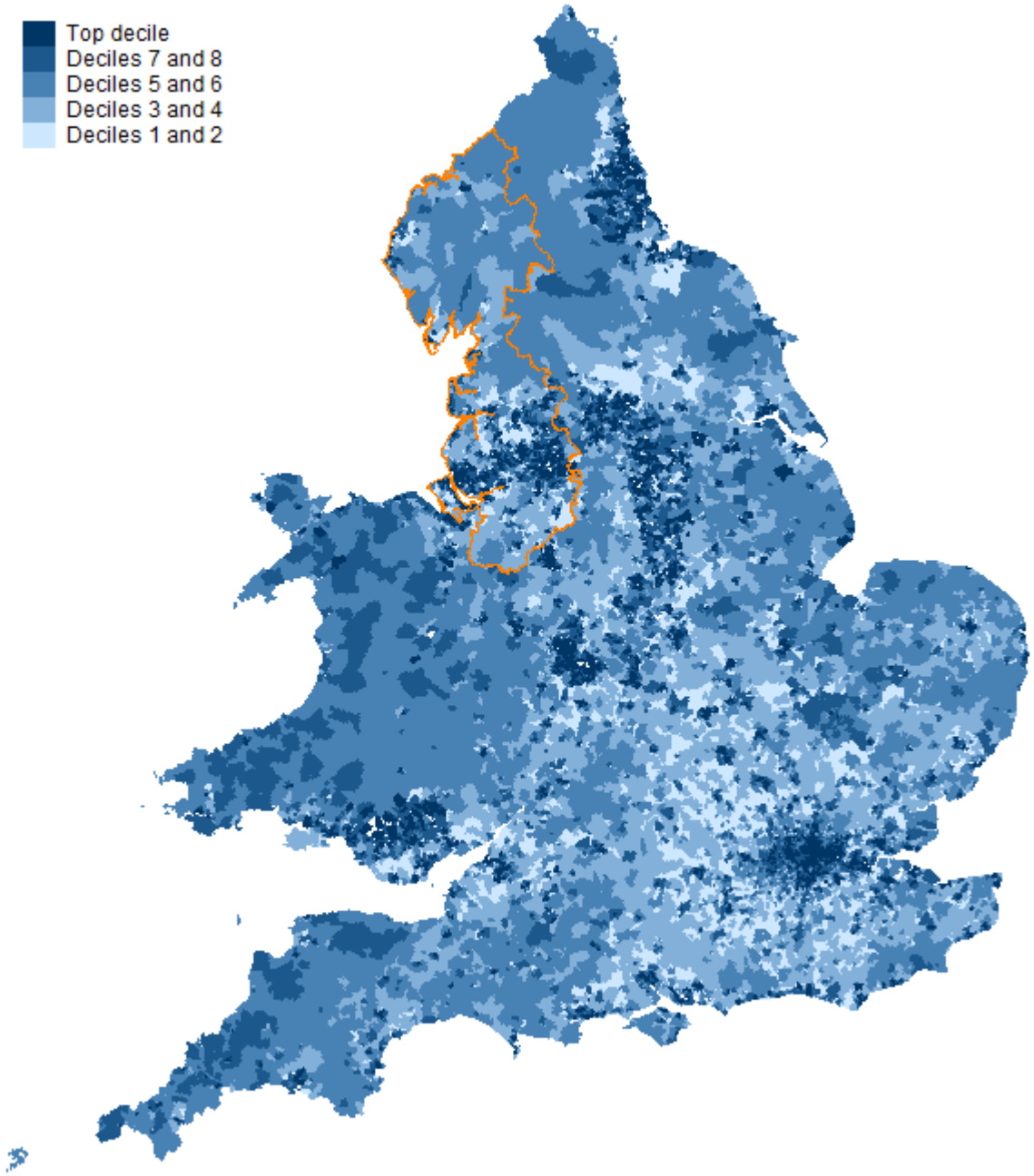


50. To illustrate the variation in the measures of the Equifax variables across LSOAs, we set out in Figure 3 a mapping of the variable "RGC100 – Postcode Risk Navigator Base - Credit Risk score derived from non- Insight data" across LSOAs. For the purpose of drawing the map we have chosen to colour the LSOAs according to the decile they fall

[9] These six variables are: "LPCF18 – Full Insight Postcode Event - % households with 1 or more Credit/Store Card accounts"; "LPCF57 – Full Insight Postcode Event - % households with total credit limit on active revolving Insight > £10,000"; LPCF72 – Insight Postcode Event - % households with worst status in last 6 months active revolving Insight = 0"; "MGC140 – Landscape Risk – Non Insight Credit Risk propensity score", "RGC100 – Postcode Risk Navigator Base - Credit Risk score derived from non-Insight data"; "RGC102 – Postcode Risk Navigator Full - Credit Risk score derived from all Insight data".

within in terms of their value for that variable. Darker shades of blue indicate that an LSOA has a value associated with higher arrears risk.

Figure 3 Map of Equifax variable "RGC100 – Postcode Risk Navigator Base - Credit Risk score derived from non- Insight data" across LSOA s (2015)



Analysis of United Utilities' debt costs at the local level

51. This section describes our analysis of how variations in the ONS deprivation measures and in the Equifax variables at the LSOA-level can explain variations in the levels of United Utilities' bad debt costs across the LSOAs within its area of appointment.

52. This analysis is focused on the area served by United Utilities because this is the only area for which data on bad debt costs at the LSOA level was available to us. Similar analysis could be carried out for the area served by other water companies if that data were to become available.

Overview of data on United Utilities' retail costs

53. United Utilities provided us with a dataset containing a breakdown of its household retail costs for each of the Lower Layer Super Output Areas (LSOAs) it serves. The data is for 2015/16 and cover 4,511 LSOAs. For each LSOA, the costs are broken down into three categories, "Doubtful debt", "Debt management (including charitable trust)" and "Other costs". Table 7 shows the breakdown of the costs across these categories. As reported in the table, there is an amount that United Utilities did not allocate between LSOAs.

Table 7 Breakdown of United Utilities' household retail costs (2015/16)

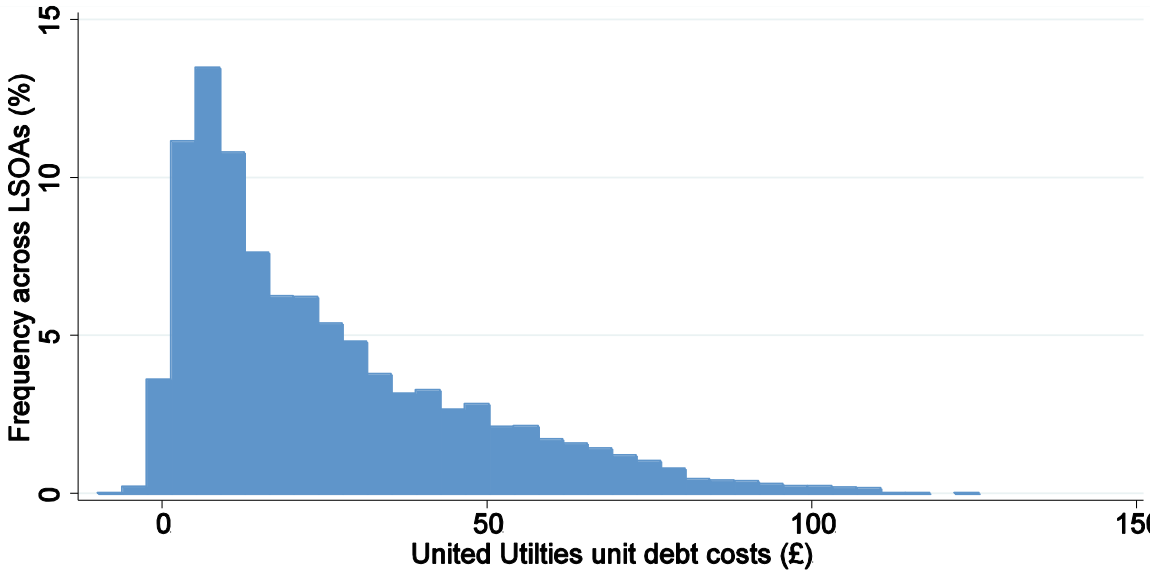
Item	Allocated across LSOAs	Unallocated
Debt management (including charitable trust)	£20.8 million	£1.2 million
Doubtful debts	£60.0 million	£0
Other	£20.8 million	£7.2 million
Total	**£101.6 million**	**£8.4 million**

54. We note that United Utilities' annual performance report for 2015/16 shows that its debt management costs for households was £11 million, lower than the £20.8 million reported in the table. United Utilities told us the additional £9.8m of debt management costs relates to donations made to the UU Charitable Trust (which were reported as part of customer service costs in the 2015/16 annual performance report) and a proportion of general support costs (recorded under Other operating expenditure in the 2015/16 annual performance report).

55. We focused our analysis on the costs associated with debt, constructed as the sum of debt management costs and doubtful debts. We think it unlikely that the quantum of costs relating to debt that were not allocated between LSOAs would distort those results. The unallocated costs relating to debt management and doubtful debt represent around 1.5 per cent of total debt management and doubtful debt costs.

56. The data provided by United Utilities includes information on the number of domestic customers in each of the LSOAs. We have used this to compute costs on a per unique customer basis. Excluding the unallocated set of costs discussed above, United Utilities' debt costs averaged at just over £26 per domestic customer, and its total retail costs were on average around £32.8 per domestic customer. There is, however, considerable variation in the household unit debt costs across the LSOAs served by United Utilities, as shown in the histogram set out in Figure 4.

Figure 4 Histogram of United Utilities' debt cost per customer across LSOAs (2015/16)



57. To draw Figure 4 we excluded the observations for two LSOAs for which the unit debt costs were large negative numbers (namely, –£443 and –£102 per customer).[10] This was for presentational reasons alone; including those two LSOAs would have stretched the axis to cover the two large negative values, thereby compressing graphically the range of values over which all the other observations lie. Figure 4 does include the observations for 52 other LSOAs for which the debt costs are reported to be negative: across those 52 LSOAs, the average debt cost is just above – £1.5 per customer.

Association between United Utilities' debt costs and ONS measures of deprivation

58. We have examined the association between United Utilities' unit debt costs and the ONS measures of deprivation.

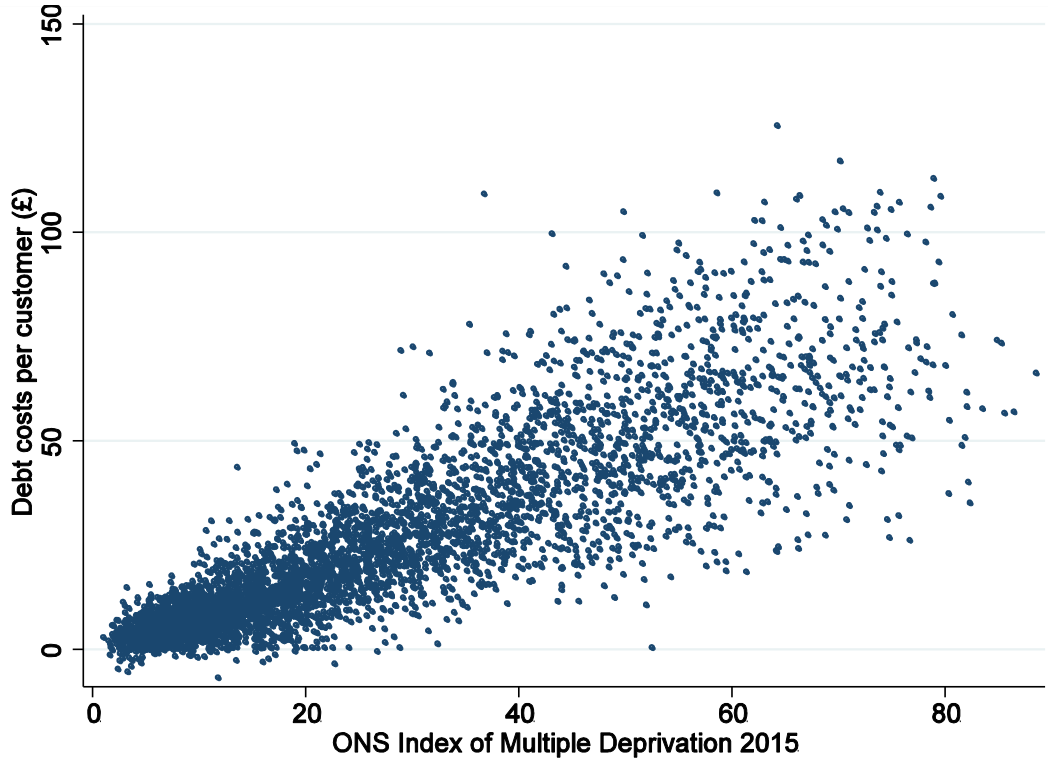
59. Figure 5 (overleaf) charts, for each LSOA in United Utilities' region, the cost of debt per domestic customer against the 2015 ONS Index of Multiple Deprivation. The figure excludes observations for six LSOAs served by United Utilities which are within Wales (and for which an ONS Index of Multiple Deprivation is not calculated).

60. Figure 5 suggests a positive association across LSOAs between the water debt unit costs of United Utilities and the ONS IMD measure of deprivation. We have explored this

[10] United Utilities told us that these negative values came about because it had been able to collect against outstanding debts that had previously been provided for in its doubtful debt charges.

further by estimating an econometric model of debt unit costs against the IMD, as well as against the other measures of deprivation calculated by ONS.

Figure 5 Household debt cost per customer against ONS IMD (2015) across LSOAs



61. Table 8 (overleaf) shows the results of that analysis for three models. The table reports the estimated coefficients and, in brackets, the t-statistic which is the ratio of the estimated coefficient to the estimated standard error of that coefficient. The results echo what is observed in Figure 6; the variation in the measures of deprivation across LSOA explain a good deal of the variation in the unit cost of debt across LSOAs.

62. We also estimated a model that included both the ONS income deprivation score and the ONS employment deprivation score. We found that the estimated coefficients for those two variables in that model were not significant. This is a reflection of the fact that the two deprivation measures are highly correlated with each other.

Table 8 Unit debt cost regressed against ONS measures of deprivation

	Model A1	Model A2	Model A3
Dependent variable	Ln (Debt costs per customer)	Ln (Debt costs per customer)	Ln (Debt costs per customer
Explanatory variables			
ONS IMD	0.046		
	(81.312)		
ONS income deprivation score		7.037	
		(83.363)	
ONS employment deprivation score			8.921
			(76.852)
Constant	1.545	1.556	1.418
	(83.708)	(86.583)	(68.279)
R-squared	0.60	0.61	0.57

63. To help interpret the regression results in Table 8, we calculated what each of the models predict would be the change in unit debt costs of serving an LSOA with a upper-quartile level of deprivation rather than one with a median level of deprivation.

64. Table 9 presents the results of these calculations. As an example, take the ONS income deprivation score. The median value of that variable across the LSOAs in United Utilities region is 13.6 per cent. The upper-quartile value for the income score is 25.3 per cent, meaning that 75 per cent of the LSOAs have an income deprivation score below 25.3 per cent. The table shows that, according to Model A2, United Utilities' unit debt costs of serving an LSOA with an income deprivation of 25.3 per cent (the upper quartile level) would be 128 per cent higher than serving an LSOA an income deprivation of 13.6 per cent (the median level). This presentation of the estimated effects of each variable attempts to capture both the size of the estimated coefficients from the regressions, and the variation in that variable across LSOAs within United Utilities' area of appointment.

Table 9 UU unit debt cost against ONS deprivation scores: implied effects

Explanatory variable	Median	75th percentile	Estimated percentage change on unit debt cost		
			Model A1	Model A2	Model A3
ONS IMD score	21.5	39.4	126%		
ONS income deprivation score	13.6%	25.3%		128%	
ONS employment deprivation score	12.4%	21.3%			121%

65. As shown in Table 9, all three models estimate that the expected effect of moving from an LSOA with the median value of a deprivation measure to one that is on the 75th percentile is associated with increasing the unit debt costs by 121 to 128 per cent. The finding that the size of the effect is of a similar size in each of the models is not surprising given that the three measures are highly correlated.

Association between United Utilities' debt costs and Equifax variables

66. We explored a range of econometric models where we regressed United Utilities' unit debt costs against variables in the Equifax dataset. We sought to develop three types of models:

(a) models where the set of explanatory variables relate to the underlying socio-economic and demographic characteristics of the LSOAs in the Equifax datasets, and to the ONS income deprivation score;

(b) models where the set of explanatory variables relate to the measures of arrears risk, as derived by Equifax; and

(c) models where the set of explanatory variables are drawn across from both categories of Equifax variables.

67. The data on United Utilities' unit debt cost refers to 2015/16. For this analysis, we used the data from the Equifax dataset reported for 2015. The number of observations used for any one regression varies according to the set of explanatory variables included, reflecting the fact that the data on some variables is not reported for a small number of the 4,511 LSOAs covered by United Utilities' dataset.

68. Table 10 reports the results for a set of exploratory models, which we have labelled Models B1 and B2.

69. Model B1 is one where the set of explanatory variables are ones that can be interpreted as measuring arrears risk. The explanatory variables in Model B2 are ones that relate to underlying socio-economic and demographic characteristics of the local area.

Table 10 United Utilities unit debt cost against Equifax variables

	Model B1	Model B2
Dependent variable	Ln (Debt costs per customer)	Ln (Debt costs per customer)
Explanatory variables		
RGC102	-0.023	
	(-23.285)	
Log of XPCF2	0.977	
	(30.966)	
AGC300		0.046
		(15.992)
GCG609		0.043
		(8.014)
GCG689		0.021
		(17.439)
GCG698		0.006
		(4.974)
ONS income score		1.508
		(5.721)
Constant	4.894	1.139
	(31.49)	(41.51)
R-squared	0.731	0.657
Number of observations	4,456	4,451

Full name of Equifax variables included in reported models

RGC102 – Postcode Risk Navigator Full - Credit Risk score derived from all Insight data

XPCF2 – Partial Insight Postcode Event – Average number of Partial Insight accounts or CCJs per household

AGC300 – Wealth Indicator - semi-decile ranking of Wealth of Postcode (1 = High Wealth, 20 = Low Wealth)

GCG609 – CENSUS Household Dependant Kids and Employment Dependent Children in Household and 0 Adults in Employment

GCG689 – CENSUS Household Car Usage 0

GCG698 – CENSUS Household Tenure - Rented LA

70. Some comments on the set of models presented in the table, as well as from the process of developing those models from a wider set:

(a) Equifax variables contribute to explaining the variation in unit debt costs across LSOAs. They add to what can be explained by controlling only for the ONS measures of deprivation. In Model B2, there is a role for a several Equifax variables, in addition to the ONS income deprivation score.

(b) We find that models that include in the set of explanatory variables ones that relate to Equifax measures of arrears risk tend to fit the data better than those where all of the explanatory variables relate to underlying socio-economic and demographic characteristics. The comparison of the fit of Model B1 and Model B2, as measured by the R-squared, is typical of that.

(c) Whilst the various Equifax measures relating to arrears risk capture different features, we found that including several of them within the same model tended to produce results that were not very robust. That is to say, the sign of the coefficient would flip or the size of the estimated coefficient change markedly. We expect this is likely to be driven by the relatively high correlation between some of those measures of arrears risk.

(d) In exploring models where the explanatory variables are ones that relate to socio-economic and demographic factors, we found that the inclusion of the ONS income deprivation score contributed significantly to the fit of the model. Where we included this ONS measure, we found that the role of the Equifax variables relating to unemployment (GCG552 and GCG557) was greatly diminished. In broad terms, the contribution of those two Equifax variables is taken up by the income deprivation measure.

(e) One consistent finding in the models we explored to date is that the Equifax variable relating to transiency, the variable labelled EPCF27 and defined as the average number of occupancy changes per household, made no obvious contribution to the models.

(f) In our analysis so far, we have not found a satisfactory model that combined Equifax variables relating to arrears risk with ones relating to the socio-economics and demographic characteristics. As above, those models that combined those categories of variables tended to produce estimates of coefficients for some of the explanatory variables included that were not intuitive.

71. Further to considering models of unit debt costs, we explored models that sought to explain variations in total household retail operating expenditure. As with the models of unit debt costs, we found that measures of deprivation — whether these were drawn from the ONS or from the Equifax dataset — explained a significant amount of the variation in the unit operating expenditure across the LSOA's served by United Utilities. Indeed, we found that, in comparison with the models of unit debit costs, the models tended to fit the data better.

Analysis of association between ONS deprivation measures and Equifax variables

72. This section reports on our work to date to develop econometric models to explain the variation in the ONS 2015 measures of deprivation by drawing on the variables from the Equifax dataset.

73. The results from the econometric models can be used to construct "predicted" values of the ONS measures for the set of Welsh LSOAs, and for the years since the ONS last published its deprivation measures. Such predicted values can then be used as candidate explanatory variables in the modelling of comparing company-level debt costs across the water sector.

IMD and income deprivation measure

74. Both the IMD and the income deprivation score are numbers that, by construction, are constrained to be between 0 and 100. Because of this, and because for both measures there are a significant number of observations at the lower end of the range, close to 0, it is better to depart from the standard ordinary least squares (OLS) to estimating the coefficients of the specified models.

75. One approach that is appropriate for such a setting is to estimate the model using "beta regressions". This is a maximum likelihood approach that carries out the estimation on a transformation of the dependent variable. Using beta regressions ensures that the predicted values obtained from the estimation of the model remain between 0 and 100.

76. The ONS 2015 measures of deprivation are based on data drawn mainly from 2012/13. To take account of this, in our analysis we have used 2012 data on the Equifax variables. An exception to this concerns the set of Equifax variables whose values were not updated from either 2006 or from 2008 through to 2013. We considered that for those variables it was reasonable to assume that the values reported for 2014 were more likely to be closer to the "true" values for 2012, than if we had used the Equifax values reported for 2012 (as these had been unchanged in most cases since 2006, or in the case of two variables since 2008).

77. Taking the IMD and the income deprivation score in turn, we explored a number of models, varying the Equifax variables included as explanatory variables. We also considered variations in the assumption regarding the function used to transform the dependent variable for the purpose of estimating the beta regression. We outline the main findings from our analysis on this so far:

    (a) We found a number of alternative candidate models that provide a good fit to the data, both with respect to IMD and in respect to the income deprivation measure.

    (b) Compared to the models discussed earlier relating to our analysis of United Utilities cost data, we found that the models developed here tended to be more stable. That is to say, the magnitude and sign of the estimated coefficients of the Equifax variables included as explanatory variables were less susceptible to changing significantly with the addition or exclusion of other variables.

(c) Also in contrast to the analysis on United Utilities' costs, we found candidate models that performed well where the explanatory variables included both variables relating to arrears risk and variables relating to socio-economic and demographic characteristics.

78. A possible explanation for the last two points lies in the fact that, for this exercise, we drew on a considerably larger dataset than was the case with the earlier analysis, almost 33,000 observations relating to the LSOAs across England, compared to the 4,500 relating to the LSOAs served by United Utilities.

79. Table 11 lists the set of Equifax variables included in the models we opted to take forward to construct predicted values of the ONS deprivation measures. We chose to use the same set of explanatory variables, and the same assumptions regarding the functional form related to the transformation function used for the beta regression, for the model where the dependent variable was the ONS' IMD and for the model where the dependent was the ONS income deprivation score. Our choice of the models to take forward to predict the ONS deprivation measures was made on the basis of our judgement concerning the role of the Equifax variables and how this tallied with the magnitude, sign and estimated variance of the estimated coefficients, and on how well the model fitted the data (based on the Bayesian Information Criteria).

Table 11 Models of ONS IMD and income deprivation score against Equifax variables

	Model C1	Model C2
Dependent variable	ONS IMD	ONS income score
Explanatory variables (common to both models)		
	LPCF72 – Insight Postcode Event - % households with worst status in last 6 months active revolving Insight = 0	
	RGC102 – Postcode Risk Navigator Full - Credit Risk score derived from all Insight data	
	Log of XPCF2 – Partial Insight Postcode Event – Average number of Partial Insight accounts or CCJs per household	
	GCG543 – CENSUS Population Qualifications None	
	GCG557 – CENSUS Population Employment Inactive Sick	
	MGC191 – Landscape Property CT A – % of households in postcode that are Council Tax Band A	
Assumption on beta regression model		
Link function	Logit	
Scale-link function	Log	
Number of observations	32,708	32,708
R-squared	0.922	0.902

80. As discussed earlier, we found there were a number of other alternative candidate models we could have taken forward to estimate the predicted values of the two ONS

deprivation measures. We expect that these would yield predicted values similar to those of Models C1 and C2.

Employment deprivation measure

81. We took a difference approach to construct predicted measures of the ONS employment deprivation score. In particular, for the purpose of our current analysis, we "predicted" the employment deprivation scores as the sum of two variables reported within the Equifax dataset:

    (a) "GCG552 – GCENSUS Population Employment Unemployed"; and

    (b) "GCG557 – CENSUS Population Employment Inactive Sick"

82. The sum of these two measures is very highly correlated with the ONS employment deprivation measure; the correlation coefficient is 0.958.

83. There is room to improve on this in future work as we are aware of one important limitation of this approach: the Equifax data on the variables GCG552 and GCG557 are not updated on an annual basis. Over the period 2006 to 2015, the Equifax dataset suggests that the values for these variables were updated only once, from 2013 to 2014.

84. In the analysis we carried out for this phase of the work we did not prioritise exploring company-level models that included the employment deprivation measure. If the role of this variable is to be more carefully explored in future, it may be desirable to examine whether the above shortcoming can be addressed. For example, it is possible that the Department for Work and Pensions or HMRC release on an annual basis data that could allow us to derive a measure of the employment deprivation indicator.

Modelling company-level retail costs

85. This section outlines our work to date in exploring econometric models that use the Equifax data to explain the variations in water companies' household debt costs and total household retail costs. We describe first the variables we have used as the dependent variables in the models and those that we considered as candidate explanatory variables. We report then on a set of models we have estimated.

Dependent variable

Dependent variables considered

86. We have developed a set of econometric models to seek to benchmark measures of water companies' expenditure relating to the provision of retail services to households. We have considered two separate measures of expenditure:

    (a) debt costs, constructed as the sum of debt management costs and doubtful debt, relating to households; and

    (b) retail costs relating to households.

87. For each of these measures of cost, we have explored two alternative formulations.

    (a) **Cost per customer (unit cost)**. We have explored models where the dependent variable is the cost per customer. This construction has a natural interpretation in a benchmarking model, as the measure of expenditure modelled is expenditure per measure of output. In this instance, the output is serving household customers. At PR14, Ofwat estimated an average cost to serve across the industry. This was based on averaging the unit cost to serve across companies. Developing econometric models of unit cost is in that same vein.

    (b) **Debt cost as a share of revenue**. We have explored models where the dependent variable is expressed in terms of the share of the household wholesale and retail revenue that it represents.

Deriving a "per customer" measure of cost

88. For the purpose of constructing measures of cost on a per customer basis, the first formulation of the dependent variable outlined above, the question arises on how customers should be counted. One issue relates to the fact that customers differ with respect to the set of services they enjoy. For the purpose of calculating the cost per customer, should customers who receive both water and sewerage services count as much as those who are connected either to only water or to only the sewerage network? We are aware that in past analyses, different approaches have been followed to deal with this:

    (a) At PR14, Ofwat calculated the number of "Unique customers (adjusted for economies of scope)" for each company, and it used this in the denominator to calculate the (unmeasured) cost to serve of each company. Ofwat constructed the measure of "Unique customers (adjusted for economies of scope)" as the sum of

[Water only customers] and [Wastewater only customers] and 1.3 * [Water and wastewater customers]. The 1.3 factor is a "specific industry adjustment to account for the economies of scope benefits associated with providing both water and wastewater household retail rather than separate water and wastewater retail services."[11] Ofwat stated the 1.3 factor was based on analysis it had carried out.

(b) In work carried out for South West Water as part of preparation for PR14, we understand Oxera constructed measures of unit expenditure using the number of unique customers as the denominator.

89. For our current analysis, we have taken this second approach. That is to say, we calculated the average cost per unique customer and used this as the dependent variable in some of the models. In those models, we have then taken account of dual service issues through the explanatory variables. In particular, in models where the dependent variable is the unit cost we have included in the set of explanatory variables a measure of the average bill.

90. The potential need for the econometric model to control for differences between companies in the average bill arises when the dependant variable is expressed on a costs per customer basis. It does not arise if the dependent variable is debt cost as a share of revenue. In those cases, controlling for variation in average bills is done through the construction of the dependent variable itself. This then has the advantage of "using up" one fewer explanatory variable than models where the dependent variable is a unit cost measure. This is welcome in the context where some of the models will be estimated on the basis of data from 18 companies. That gain, however, may come at a cost. In particular, in models where the dependent variable is the debt cost as a share of household revenue there is an implicit constraint that the debt costs varies one to one with variation in household revenue. In our analysis, we may be able to observe from the models estimated whether or not such a constraint is a reasonable one to make.

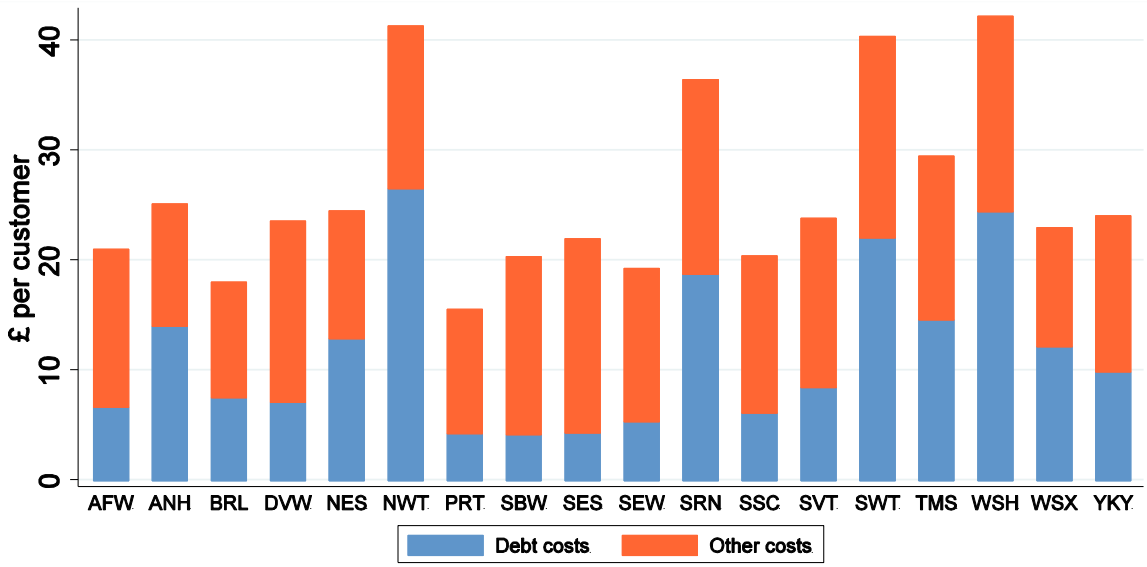
Comparison of retail costs across water companies

91. We compiled a dataset on companies' retail costs associated with serving household by extracting information from companies' annual accounts for the years in the period from 2013/14 to 2015/16. United Utilities told us that retail cost data prior to 2013/14 may be less consistent across companies.

92. Figure 6 shows the debt costs and total retail operating costs, expressed on a "per unique customer" basis, across the 18 water companies, in the period 2013/14 to 2015/16.

[11] Ofwat (2014) "Final price control determination notice: policy chapter A5 – household retail costs and revenues", Annex 1, page 35.

Figure 6 Debt cost and retail operating cost per customer (average 2013/14 to 2015/16)



Explanatory variables

93. We describe below the set of explanatory variables we considered in the models explored.

Explanatory variables derived from Equifax dataset

94. The set of explanatory variables we have considered include:

    (a) The variables included in the Equifax dataset.

    (b) The "predicted" values of the ONS IMD, income deprivation score and employment deprivation score for each water company. We calculated these on the basis of the econometric analysis described earlier, drawing on the Equifax dataset.

    (c) Variables constructed as the proportion of households served by a company who live in LSOAs that are within the 10 per cent most deprived LSOAs across England and Wales, drawing on the "predicted" values of the ONS measures of deprivation. We also constructed analogous variables for the 20 per cent most deprived LSOAs.

95. The variables described in the first two bullet points above are intended to capture average levels of deprivation or arrears risk across the area served by a company. In contrast, the variables described in point (c) are measures of the incidence of more extreme deprivation within each company's region.

96. The variables in the Equifax dataset and of our "prediction" of the ONS measures of deprivation are at the LSOA level. It was necessary for us to aggregate each of these variables to the company-level. A challenge of producing such company-level

aggregates of these variables comes from the fact that water and sewerage companies offer different services over different parts of the area they serve. How should the deprivation measures or values of the Equifax variables be aggregated across the LSOAs when the water service area of appointment differs from the sewerage service area of appointment? Possible approaches include:

(a) Calculate the weighted average of all LSOAs within a company's water supply area, with the weights given by household numbers (or by population, depending on the deprivation measure) in each LSOA.

(b) Calculate the weighted average of all LSOAs where a company provides water or wastewater or both services, with the weights given by household numbers (or by population, depending on the deprivation measure) in each LSOA.

(c) As with (b), but use as weights the product of household numbers (or of population, depending on the deprivation measure) and average bill.

97. Under approach (a), a company's average deprivation measure would not be affected by the deprivation in those LSOAs where it only provides wastewater services. Under approach (b), the deprivation levels in LSOAs where a company provides water services will contribute just as much as those where it provides wastewater services, or as much as those where it provides both services, assuming equal number of households or of population. Finally, under approach (c), the weight of each LSOA reflects both the size of that LSOA (in terms of households or population) and the average bill. This implies that the contribution of an LSOA's deprivation measure to the company-wide average is greater if the company provides both water and wastewater services to that LSOA than if it provides only one of those services.

98. The choice of approach can make a difference. Table 12 overleaf compares the weighted average ONS income deprivation score (not our fitted estimate of that measure) for each water and wastewater company (other than Welsh Water) under the three approaches.

99. As expected, the table shows that the choice of approach has little effect on the company-level measures for those companies whose water supply area and sewerage service areas are largely overlapping. That is the case for United Utilities. For other companies, such as Northumbrian Water, Wessex Water and Thames Water the differences are greater. In the case of Thames Water, the deprivation measure is greater if the water supply area alone is taken into account; the opposite is the case for the other two companies.

100. We have used approach (c) in our analysis. This seems preferable to either of the other two approaches. It does not leave out the contribution of LSOAs where a company only provides wastewater services, and the weights used reflect the relative contribution of different LSOAs to a company's revenue.

Table 12 Comparison of company-level aggregations of ONS income deprivation score

Company	Approach (a)	Approach (b)	Approach (c)
ANH	0.128	0.127	0.126
NES	0.170	0.172	0.175
NWT	0.174	0.173	0.173
SRN	0.132	0.119	0.122
SVT	0.154	0.157	0.156
SWT	0.134	0.134	0.134
TMS	0.150	0.137	0.143
WSH	—	—	—
WSX	0.102	0.114	0.110
YKY	0.161	0.160	0.161

Other explanatory variables

101. We considered other candidate explanatory variables:

(a) Average revenue per unique customer, relating to households, which is a measure of the average household bill.

(b) Proportion of unique customers who are dual service customers.

102. We compiled data on these variables drawing on the information published by Ofwat for the retail review at PR14 and on companies' annual accounts and Annual Performance Report.

103. With regard to the data on revenue per unique customer for South West Water, we applied a £50 deduction for the years from 2013/14 (inclusive) to reflect the Government's contribution to households in respect of their water bills.

Model dynamics

104. We explored two different approaches to model dynamics:

(a) We "averaged" data on the dependent variable and on the explanatory variables over time, and used the averaged values in the estimation of the models. Under this approach, for each variable, we used the average value over the three-year period, 2012/13 to 2015/16.

(b) We estimated models drawing on the value of the dependent and explanatory variables in each of the three years over the period 2012/13 to 2015/16.

105. In the set of models presented in this paper we have focused on the first approach.

106. We understand the Equifax data is reported on the basis of a calendar year. In contrast, the companies' cost data covers financial years. For the purpose of our analysis, we have mapped the Equifax data for year 201*X* with company cost data for the financial year 201*X/[X+1]*.

Modelling results

107. This section sets out and discusses the results for a set of models we explored.

Estimated models

108. Tables 13 and 14 shows the results for a set of models, which differ in terms of the choice of dependant variable and the choice of explanatory variables.

Table 13 Company-wide models of debt costs: Models D1 – D4

	Model D1	Model D2	Model D3	Model D4
Dependent variable	Log of unit debt cost	Log of unit debt cost	Ratio of debt costs to revenue	Ratio of debt costs to revenue
Explanatory variable				
Log of revenue per customer	1.102	1.108		
	(7.256)	(7.655)		
Share of households in top decile of ONS Income Deprivation (Predicted)	1.803		0.092	
	(1.674)		(2.337)	
Share of households in top quintile of ONS Income Deprivation (Predicted)		1.357		0.065
		(1.98)		(2.532)
Constant	-3.922	-4.034	0.034	0.031
	(-4.814)	(-5.192)	(8.478)	(6.272)
R-squared	0.824	0.834	0.254	0.286
Observations	18	18	18	18

109. Table 13 reports on a set of models where the explanatory variables used relates to the share of households within a company's area which live in LSOAs that are ranked within the top decile or quintile of deprivation, as measured by the predicted ONS Income Deprivation score, across all LSOAs in England and Wales. For the models in

Table 14, we have used as explanatory variables ones that capture the "average" level of deprivation across the region served by each company. We discussed the derivation of these variables earlier in the report.

110. With regard to the choice of dependent variable, we considered two alternatives. For models D1 and D2 and models E1, E2 and E3 the dependent variable is the natural logarithm of unit debt costs as the dependent variable. For models D3 and D4, and for models E4 and E5 the dependent variable is the ratio of debt costs to the company revenue from households.

Table 14 Company-wide models of debt costs: Models E1 – E5

	Model E1	Model E2	Model E3	Model E4	Model E5
Dependent variable	Log of unit debt cost	Log of unit debt cost	Log of unit debt cost	Ratio of debt costs to revenue	Ratio of debt costs to revenue
Explanatory variable					
ONS IMD score (predicted)	3.056			0.139	
	(2.297)			(2.832)	
ONS income score (predicted)		4.547			
		(2.386)			
Log of revenue per customer	1.076	1.083	1.086		
	(7.542)	(7.741)	(7.633)		
RGC102			-0.035		-0.003
			(-2.254)		(-3.06)
GCG698					-0.002
					(-1.754)
Constant	-4.286	-4.323	1.1	0.012	0.468
	(-5.745)	(-5.849)	(0.444)	(1.117)	(3.288)
R-squared	0.845	0.848	0.844	0.334	0.428
Observations	18	18	18	18	18

Full name of Equifax variables included in reported models

RGC102 – Postcode Risk Navigator Full - Credit Risk score derived from all Insight data
GCG698 – CENSUS Household Tenure - Rented LA

Diagnostic tests

111. We carried out a series of diagnostic tests on the models reported above. These are tests that PwC reported on in their reviews of the econometric models of doubtful debt costs which some companies put forward at PR14.[12]

(a) The Ramsey RESET and the Linktest models are two different model specification tests; in brief, they test for the significance of powers of the fitted values of the dependent variable were these to be included within the set of explanatory variables.

(b) The Breusch-Pagan is a test for heteroscedasticity.

(c) The Shapiro-Wilk a test of whether the residuals are normally distributed.

112. In each of these tests, the null hypothesis is that the assumption underpinning the ordinary least square estimation and inference holds e.g. in the case of the Ramsey RESET test the null hypothesis is that the estimated coefficients of the powers of the fitted values of the dependent variables are 0, and in the case of the Breusch-Pagan test the null hypothesis is that the variance is not a function of the fitted values.

113. Tables 15 and 16 tables report the outcome of the tests. The tables report the significance level of each of the tests.[13] A high value is indicative that the null hypothesis cannot be rejected, for example in the case of the Ramsey RESET that we cannot reject the hypothesis that the estimated coefficients of powers of the fitted variable are 0, or in the case of the Breusch-Pagan that we cannot reject the hypothesis that the variance is not a function of the fitted values. For ease of interpretation, we have used a 5 per cent significance level as the threshold to colour in the cells in either pale green or in amber.

Table 15 Summary of diagnostic tests: Models D1 – D4

	Model D1	Model D2	Model D3	Model D4
Ramsey RESET test for model specification	0.449	0.265	0.006	0.004
Linktest model specification test	0.275	0.299	0.710	0.654
Breusch-Pagan test for heteroscedasticity	0.574	0.383	0.270	0.136
Shapiro-Wilk test for normality of residuals	0.485	0.546	0.132	0.270

[12] For example, PwC reviews of the econometric models present by South West Water (29 April 2014), by Northumbrian Water (29 August 2014), by Welsh Water 912 December 2014) and by Thames Water (12 December 2014).

[13] In the case of the Linktest model specification test, the significance level reported is that of the test of the null hypothesis that the coefficient on the square of the predicted values is 0.

Table 16 Summary of diagnostic tests: Models E1 – E5

	Model E1	Model E2	Model E3	Model E4	Model E5
Ramsey RESET test for model specification	0.148	0.208	0.258	0.024	0.755
Linktest model specification test	0.477	0.465	0.429	0.771	0.664
Breusch-Pagan test for heteroscedasticity	0.291	0.296	0.151	0.118	0.327
Shapiro-Wilk test for normality of residuals	0.350	0.172	0.089	0.272	0.801

Comment on modelling results

114. There are a number of interesting observations to bring out from the set of results tabled, and from the analysis of other alternative models that we considered.

115. In models where the dependent variable is unit debt cost, there is a big role played by the variable controlling for average revenue per customer. The estimated coefficient for this variable tended to be close to 1, suggesting that an *X* percent change in the average bill is associated with a change in unit debt costs also of *X* percent. Finding that the estimated coefficient tended to be close to 1 suggests that the restriction implicit in the set of models that use ratio of debt costs to revenue may be a reasonable one to make (Models D3, D4, E4 and E5).

116. The set of results from Models D1 to D4 show that variations in measures of "extreme" deprivation contribute to explaining variations in unit costs, and variations in ratio of bad debt costs to revenues.

117. As shown in the results for Models E1 and E2, we found that variations in the weighted average "predicted" ONS measures of deprivation contributed to explaining variation in unit costs. We found that including the IMD or the income measure of deprivation was preferable to including the predicted employment deprivation score.

118. In models of unit debt cost, we found that using measures of arrears risk to control for deprivation produced results similar to those that included predicted ONS measures. In the table above, we report under model E3 the results of a model that includes the variable RGC102, an Equifax proprietary measure described as "Postcode Risk Navigator Full – Credit Risk score derived from all Insight data".

119. As illustrated by the results for Models D3, D4, E4 and D5, the models we explored were able to explain a smaller share of the variation in the cost measure when the cost measure used as the dependent variable is the ratio of debt costs to revenue. The result is not surprising. In essence, it reflects the fact that the average size of bill is an important driver in the comparison of unit debt across companies. In the case of Models D3, D4, E4 and E5, that effect is already taken account of in the construction of the dependent variable, and the explanatory variables are seeking to control for variations

relating to other factors. The reported R-squared is not directly comparable across models where the dependent variable differs.

120. We found that in models that included more than one measure controlling for deprivation, the estimated coefficients on the two variables tended to be counter-intuitive and not statistically significant. We expect this is a consequence of the high correlation between the different measures of deprivation, and the fact we had a sample of only 18 observations on which to estimate the model. It may be possible to develop alternative and more refined model specifications that allow the incorporation of two explanatory variables relating to deprivation and/or arrears risk.

121. We also considered models where we controlled for the mix of services supplied, by including as an explanatory variable the ratio of the number of dual service customers to the number of unique customers, rather than by including the variable on revenue per customer. We found that including the ratio of dual service customers to unique customers alongside a variable relating to deprivation tended to undermine the model as a whole. For interest, we estimated models where we regressed the logarithm of unit debt costs against the ratio of dual service customers to unique customers, with no other explanatory variable. The hypothesis for such models is that unit debt costs are driven by the mix of customers, and that levels of deprivation of average bill size or other factors are not important drivers. We found that in such models the estimated coefficient on that variable was around 1.3. The interpretation of this is that there are diseconomies of scope associated with serving dual service customers, which we find to be counter-intuitive.

122. We also explored models where the dependent variable is unit total retail operating costs associated with serving households, rather than unit debt costs. We found that in these models, the estimated coefficients for the explanatory variables relating to predicted ONS measures of deprivation as well as those relating to arrears risk reported in the Equifax datasets tended to have much wider confidence intervals than in those models where the dependent variable is unit debt cost. In various models, we found those variables not to be statistically significant.

123. The set of results discussed so far relate to models where we averaged the values of the dependent and of the explanatory variables across the more recent three years of data, spanning the period 2013/14 to 2015/16. We have done some initial examination of models where we draw on data for each of those three years separately, rather than averaging them over time. We estimated those models using ordinary least squares, with robust standard errors. We ran models where we included separate dummy variables for each of the years span by our dataset, and models where we included a time trend. Our work so far point to results fully in line with the findings from the models based on data that have been averaged over the three-year period. As with the earlier models, we find that measures of deprivation and of the average bill explain much of the variation in unit debt costs.