Water Resources Management Plan 2024

Technical Appendix F – Household Demand Forecasting

Background and Introduction

For this appendix we have published a report written by Artesia Consulting which describes the approach and methods used for developing a household consumption forecast for the period 2020 through to 2100. Two separate household consumption models were developed, one for the London Water Resource Zone (WRZ), and one for the for the remaining water resources zones (collectively referred to as 'Thames Valley'). Both of these models remain unchanged to those used in WRMP19.

The Water Industry best practice for household demand forecasting was updated in 2015 and remain unchanged. We continue to use a forecasting method based on multiple linear regression (MLR) for all WRZs.

We continue to consider MLR the most appropriate method because:

- Using this method, forecasts ca be built on good quality historic household consumption, demographic and housing type data from across the Thames Water region
- The models are based on demographic and housing factors that are known to influence household consumption
- The models could be constructed using standard statistical methods that allow the robustness and uncertainties of the models to be quantified
- The models could be validated using historic and regional data

The appendix describes how models for household consumption were developed using 10 years' worth of validated historic individual household demand data from a sample of 3000 households across the Thames region. It describes the selection of the variables used to explain changes in consumption, and how these were built into the models. The approach for collecting and preparing the parameter data for the variables is described, and it is explained how these are applied to the models to predict consumption in unmeasured households and measured households.

The method of model testing and validation is also described, along with model calibration, normalisation of the base year and uplifts for dry year, critical period, and climate change scenarios. The application of these models to produce household demand forecasts to 2100 is described within Section 3: Current and future demand for water. The application of models when calculating demand uncertainty for Target Headroom is explained within Section 6 Uncertainty and Baseline Supply Demand Balance.

Artesia
Consulting

Thames Water

WRMP19 Household Consumption Forecast

Final Report (version 2.3)

AR1172

04-10-2017

Report title:	WRMP19 Household consumption forecast
Report number:	AR1172
Date:	04-10-2017
Client:	Thames Water
Author(s):	Redacted

Any enquiries relating to this report should be referred to the authors at the following address:

Artesia Consulting Ltd, Unit 2 Badminton Court, Yate, Station Road, Bristol, BS37 5HZ.

Telephone: + 44 (0) 1454 320091

Website: www.artesia-consulting.co.uk

In preparing this report, Artesia has relied, in whole or in part, on data and information provided by Thames Water (the "Client") and third parties, which information has not been independently verified by Artesia and which Artesia has assumed to be accurate, complete, reliable, and current. While Artesia has

Executive Summary

Artesia Consulting were tasked with developing and delivering a model to produce a baseline household consumption forecast, which could be projected forward to the year 2100.

Thames Water required the most appropriate forecasting method for household consumption in each WRZ. This identified that the London Water Resource Zone (WRZ) required a bespoke model due to the overarching importance of this area. Separating the Thames Valley dataset from the London modelling mitigated the risk of potentially skewing the model for the largest area (London) by the five Thames Valley water resource zones.

Within this report we have discussed how the forecasting methods have been selected, developed, applied, tested, validated and calibrated. We have also made an assessment of the overall uncertainty in the forecast.

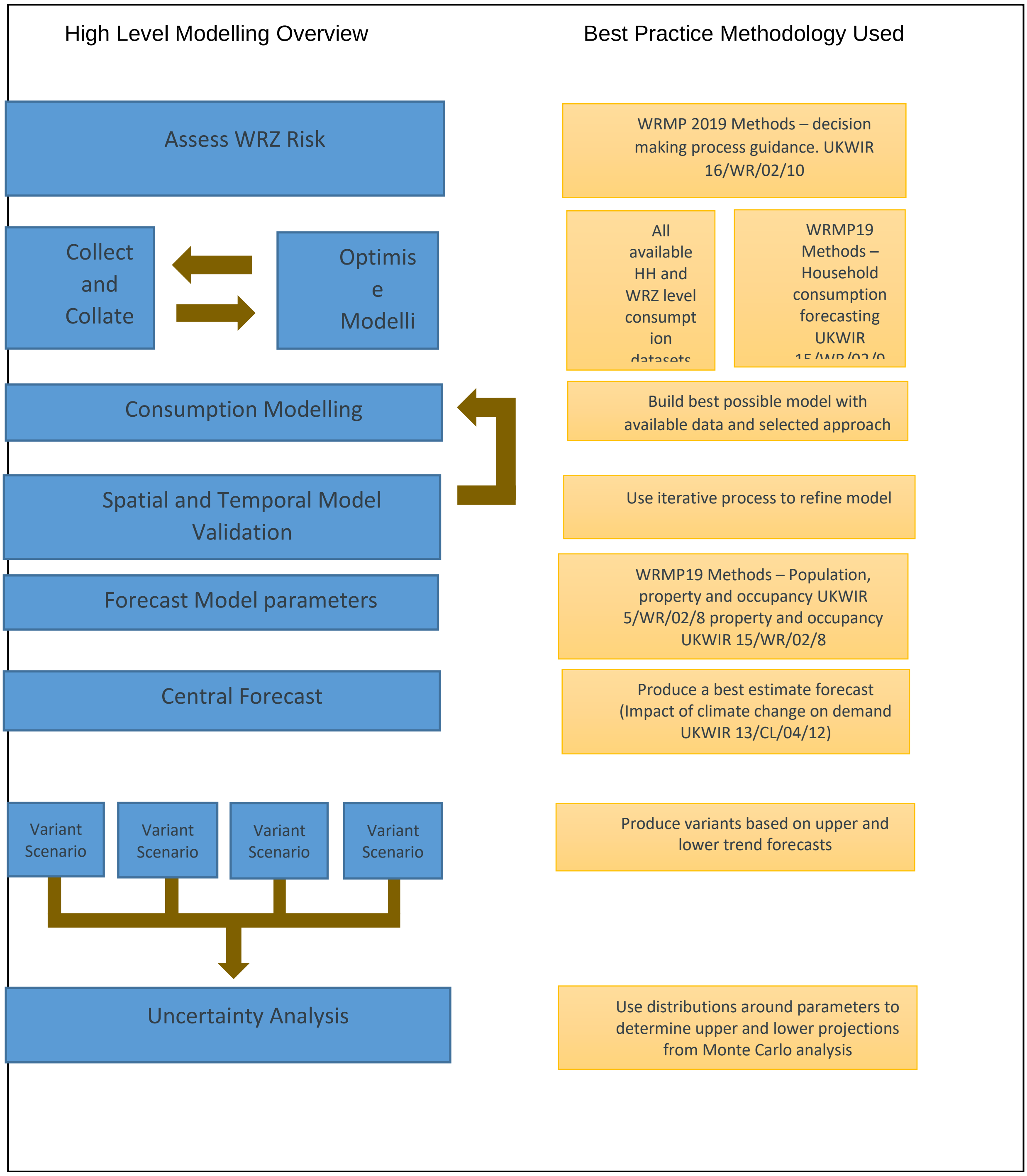
Multiple linear regression (MLR) modelling has been applied using 10 years of validated historic data to create a consumption model to predict usage in each water resource zone. These have then been used to forecast the first 25 years of the planning period. The forecasts have then been extrapolated to 2100. During the current AMP cycle, Thames Water are carrying out an extensive customer metering programme, and the potential impact from this programme has been incorporated into the model outputs, using data supplied by Thames Water.

For the previous three water resource plans Thames Water have used a micro-component modelling approach. The switch to a different methodology for modelling (the use of MLR) places a greater emphasis on validation of the new model. It is not possible to directly compare the outputs from the old micro-component model and the new approach. Therefore, within this report we have validated the MLR models using three different approaches. Firstly, the models are constructed using standard statistical methods from which the uncertainty can be quantified. Secondly, the models have been validated temporally, by applying the model to historic data and forecasting forwards to the current year and comparing with reported figures. Thirdly, the models have been validated spatially, by applying the model to about 240 sub-zones across the Thames Water region and comparing with reported data. This is a level of validation that could not be carried out with previous micro-component based models.

The implementation of the model has been designed to produce all the necessary outputs for completing the WRMP19 tables, including an estimate of the micro-component consumption derived from the household figures produced by the MLR model forecasts.

The MLR household consumption forecast has been developed using the latest industry best practice, the figure on the next page illustrates the high level modelling process and some of the key best practice that has been followed.

The resulting multiple linear regression (MLR) model is a robust model that integrates the impacts of drivers such as occupancy, property type, socio-demographics and meter penetration in a dynamic model which can be used to test sensitivities around individual parameters.



High Level Modelling Overview

Best Practice Methodology Used

Assess WRZ Risk

WRMP 2019 Methods – decision making process guidance. UKWIR 16/WR/02/10

Collect and Collate

Optimise Modelli

All available HH and WRZ level consumption datasets

WRMP19 Methods – Household consumption forecasting UKWIR 15/WR/02/9

Consumption Modelling

Build best possible model with available data and selected approach

Spatial and Temporal Model Validation

Use iterative process to refine model

Forecast Model parameters

WRMP19 Methods – Population, property and occupancy UKWIR 5/WR/02/8 property and occupancy UKWIR 15/WR/02/8

Central Forecast

Produce a best estimate forecast (Impact of climate change on demand UKWIR 13/CL/04/12)

Variant Scenario

Variant Scenario

Variant Scenario

Variant Scenario

Produce variants based on upper and lower trend forecasts

Uncertainty Analysis

Use distributions around parameters to determine upper and lower projections from Monte Carlo analysis

Contents

Tables

Figures

1 Introduction

1.1 Glossary

Throughout this report, there are many terms used which could be deemed as having multiple meanings. For clarity, the following glossary table sets out the definitions that we have used for the following terms:

Error	The deviation of the observed value from the unobservable true value of a quantity of interest.
Measured household	A household who has a water meter installed, and is billed according to their usage. Sometimes called 'metered households'.
Multiple linear regression (MLR)	An extension to simple linear regression. It is used to predict the value of a variable based on the value of two or more other variables. The variable we wish to predict is called the response variable, the variables we are using to predict are called the explanatory variables.
Outlier	An observation which lies outside of the overall pattern of a distribution.
Uncertainty	A range of values within which the true value is asserted to lie.
Unmeasured household	A household who does not have a water meter installed, and is therefore billed according to their rateable value. Sometimes called 'unmetered households'.
Residual	The difference between the observed value and the estimated value.

1.2 Context

The duty to prepare and maintain a Water Resource Management Plan (WRMP) is set out in Section 37A-37D of the Water Industry Act. This report sets out the development of the Household consumption component of the WRMP for Thames Water. As explained in the next section, the household consumption forecast has been developed using the latest industry best practice.

The final version of the Environment Agency water resources planning guidelines were published in May 2016. Section 5 of the technical guidance describes the approaches that water companies should take to forecast demand, including household consumption. In particular, section 5.4 of the guidance states that:

"You should select demand forecasting methods appropriate to the data available and the supply-demand situation in individual water resource zones".

Thames Water wish to identify the most appropriate forecasting method for household consumption in each WRZ. The method needs to address all the regulatory, business and risk drivers, and provide a sound foundation for WRMP19. Within this report we have discussed how the forecasting methods have been selected, developed, applied, tested, validated and calibrated. We have also made an assessment of the overall uncertainty in the forecast.

This report discusses the development of the baseline forecast, which has been projected forward to the year 2100. Multiple linear regression (MLR) modelling has been developed using 10 years of validated historic data to create consumption models for each water resource zone. Following the model build, it may be possible to obtain additional historic data dependent on the variables selected in the final model. This will enable further testing, and means that additional data can be included in the trainer set.

During the current AMP cycle, Thames Water are carrying out an extensive customer metering programme, and the potential impact from this programme has been incorporated within the model outputs using data provided by Thames Water.

For the previous three water resource plans Thames Water have used a micro-component modelling approach. The decision to use an MLR approach is discussed within the report. The switch to a different methodology for modelling places greater emphasis on validation of the model. It is not possible to directly compare the outputs from the old micro-component model and the new approach, therefore, within this report we have validated the MLR models using three different approaches. Firstly the model is constructed using standard statistical methods from which uncertainty can be quantified. Secondly the model has been validated temporally, by applying the model to historic data and forecasting forwards to the current year and comparing with reported figures. Thirdly, the model has been validated spatially, by applying the model to about 240 sub-zones across the Thames Water region and comparing with reported data. This is a level of validation that could not be carried out with previous micro-component based models.

The implementation of the model has been designed to produce all the necessary outputs for completing the WRMP19 tables, including an estimate of the micro-component consumption derived from the household figures produced by the MLR model forecasts.

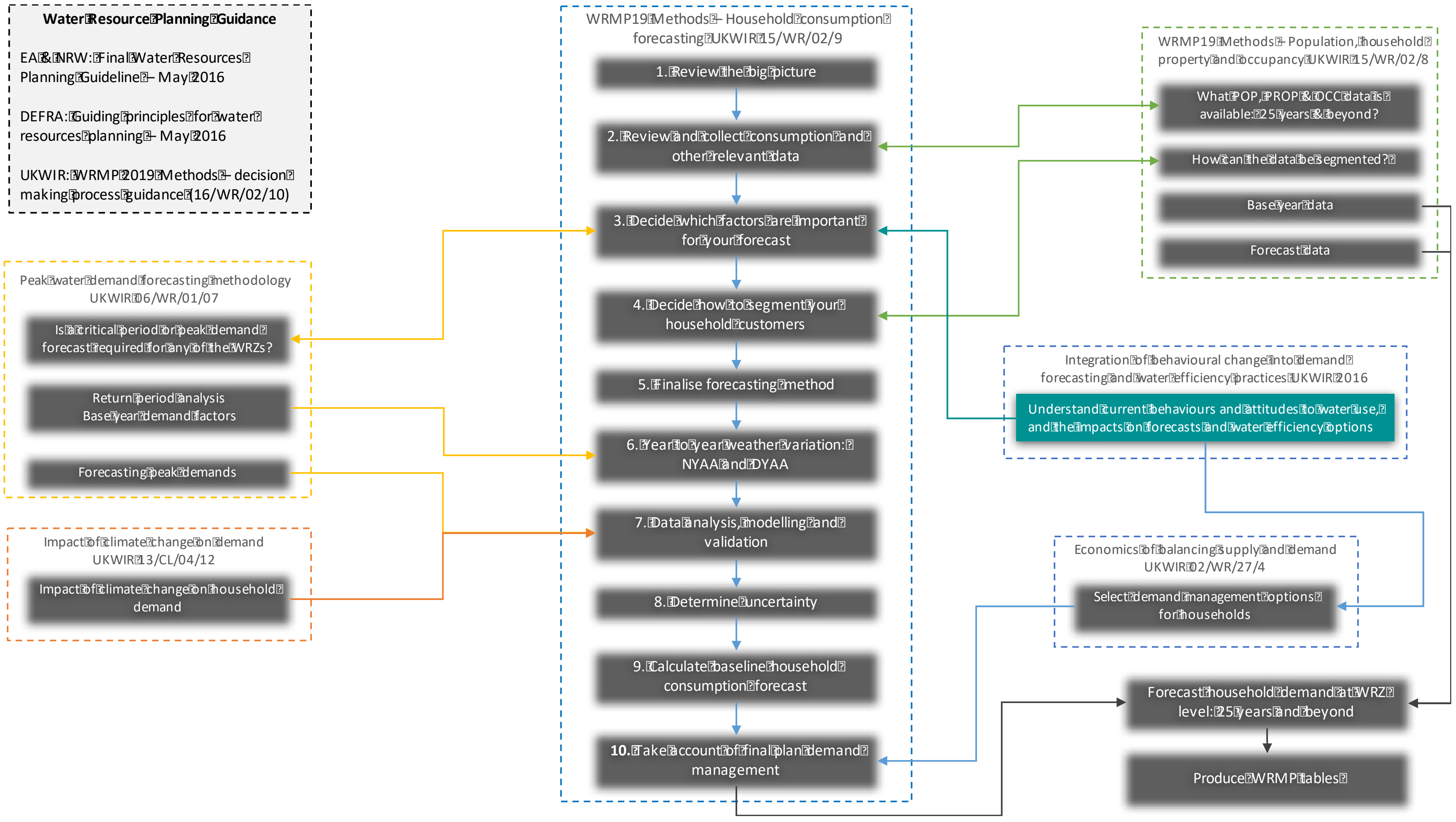
1.3 Approach and best practice

The derivation of the household consumption forecast follows UKWIR best practice guidance[1]. Figure 1 shows the overall methodology contained in this best practice in the centre blue boxes, and then linked to this are the various other best practice and guidance reports that are relevant to household consumption forecasting.

The progression through the various stages of household consumption is inevitably an iterative process and so this report groups some of the stages together in logical steps, contained in the main sections below.

[1] WRMP19 Methods – Household consumption forecasting. UKWIR. 2015. Ref: 15/WR/02/9

Figure 1 Best practice guidelines and approach


Water Resource Planning Guidance
EA & NRW: Final Water Resources Planning Guideline – May 2016
DEFRA: Guiding principles for water resources planning – May 2016
UKWIR: WRMP 2019 Methods – decision making process guidance (16/WR/02/10)
Peak water demand forecasting methodology UKWIR 06/WR/01/07
Is a critical period or peak demand forecast required for any of the WRZs?
Return period analysis Base year demand factors
Forecasting peak demands
Impact of climate change on demand UKWIR 13/CL/04/12
Impact of climate change on household demand
WRMP19 Methods – Household consumption forecasting UKWIR 15/WR/02/9
1. Review the big picture
2. Review and collect consumption and other relevant data
3. Decide which factors are important for your forecast
4. Decide how to segment your household customers
5. Finalise forecasting method
6. Year to year weather variation: NYAA and DYAA
7. Data analysis, modelling and validation
8. Determine uncertainty
9. Calculate baseline household consumption forecast
10. Take account of final plan demand management
WRMP19 Methods – Population, household property and occupancy UKWIR 15/WR/02/8
What POP, PROP & OCC data is available: 25 years & beyond?
How can the data be segmented?
Base year data
Forecast data
Integration of behavioural change into demand forecasting and water efficiency practices UKWIR 2016
Understand current behaviours and attitudes to water use, and the impacts on forecasts and water efficiency options
Economics of balancing supply and demand UKWIR 02/WR/27/4
Select demand management options for households
Forecast household demand at WRZ level: 25 years and beyond
Produce WRMP tables

2 Method Selection

The first stage is to look at the big picture (in Water Resource Planning terms), determine the planning challenges that exist, do they differ between water resource zones , what length of forecast is required, what data is available and which method or methods will be most suitable for the household consumption forecast.

The draft outputs from the UKWIR 'Decision making process' framework[2] were used to characterise the water resources planning problem in Thames Water's water resource zones (WRZs), as illustrated in Table 1.

Table 1 Problem characterisation based on WRMP14

Zone	Problem characterisation
London	High
SWOX	High
SWA	Medium
Kennet Valley	Low
Guildford	Low
Henley	Low

The UKWIR Household consumption forecasting guidance identifies the following methods for forecasting household consumption (in approximate order of complexity):

- Use existing study data;
- Trend based models;
- Per-capita methods;
- Variable flow methods;
- Macro-components (referred to as 'major consumption groups' hereafter);
- Micro-components;
- Regression models;
- Proxies of consumption; and
- Micro-simulation.

A full review of these methods is presented in the UKWIR Household consumption forecasting guidance, and further information is included in a previous report for Thames Water[3].

[2] WRMP 2019 Methods – decision making process guidance. UKWIR. 2016. Ref: 16/WR/02/10

[3] Review of methods for forecasting household consumption for Thames Water. Artesia. March 2016. Ref: AR1103.

The criteria presented in Table 2 were developed in the UKWIR consumption forecasting guidance to assess the forecasting methods.

Table 2 Criteria for evaluation consumption forecasting methods

Criteria	Comment
Acceptance by stakeholders	The method should stand up to scrutiny from the regulators, and other external stakeholders, including customers.
Explicit treatment of uncertainty	The method should recognise that there will be uncertainty around the forecast, and should quantify the level of uncertainty.
Underpinned by valid data	The method should be based on data that is valid for the area under consideration.
Transparency and clarity	The method needs to be understood and should be able to be replicated by others.
Appropriate to level of risk	The method should be appropriate in terms of cost and data requirements for the planning problem being addressed; i.e. the degree of vulnerability to a supply demand deficit.
Logical and theoretical approach	The method should command confidence to practitioners and decision makers. It should address those factors that people believe drive water demand, and it should be relevant to historical trends.
Empirical validation	The method should enable comparison to outturns or past projections. It should be possible to test the method on past data to predict demand, and predict any explanatory factors used in the forecast.
Explicit treatment of factors that explain HH consumption	The method should be able to take account of the different factors which drive household demand, and different segments of consumers with respect to household water use.
Flexibility to cope with new scenarios	The method should be method flexible enough to run different household consumption forecasts.

UKWIR best practice guidance for forecasting household consumption recommends the use of a Red Amber Green (RAG) matrix to identify the most suitable consumption forecasting methodologies.

Table 3, Table 4 and Table 5 present RAG matrix assessments of the available methods for London (high vulnerability), SWOX (medium vulnerability) and the four low vulnerability zones, respectively. These matrices have been derived, including each criteria’s weighting factor, following consultation with Thames Water.

Table 3 Matrix of forecasting methods suitable for high concern WRZs

High concern zones	Weighting	Regression models	Micro-component models	Major consumption groups	Micro-simulation	Proxies of consumption
Acceptance by stakeholders	10	8	6	7	5	2
Explicit treatment of uncertainty	5	7	5	5	7	2
Underpinned by valid data	7	7	5	6	4	2
Transparency and clarity	7	7	6	5	2	2
Appropriate to level of risk	7	10	7	10	2	2
Logical and theoretical approach	7	8	5	7	5	2
Empirical validation	5	8	5	7	5	2
Explicit treatment of factors that explain HH consumption	8	8	5	8	5	2
Flexibility to cope with new scenarios	5	7	5	7	5	2
Weighted score		**478**	**336**	**425**	**266**	**122**
Ranked		**1**	**3**	**2**	**4**	**5**

Table 4 Matrix of forecasting methods suitable for medium concern WRZs

Medium concern zones	Weighting	Regression models	Micro-component models	Major consumption groups	Variable flow methods	Trend-based models
Acceptance by stakeholders	10	8	6	7	5	2
Explicit treatment of uncertainty	5	7	5	5	7	2
Underpinned by valid data	7	7	5	6	4	2
Transparency and clarity	7	7	6	5	2	2
Appropriate to level of risk	7	9	6	10	7	2
Logical and theoretical approach	7	8	5	7	5	2
Empirical validation	5	8	5	7	5	2
Explicit treatment of factors that explain HH consumption	8	8	5	8	5	2
Flexibility to cope with new scenarios	5	7	5	7	5	2
Weighted score		**471**	**329**	**425**	**301**	**122**
Ranked		**1**	**3**	**2**	**4**	**5**

Table 5 Matrix of forecasting methods suitable for low concern WRZs

Low concern zones	Weighting	Regression models	Micro-component models	Major consumption groups	Variable flow methods	Trend based models	Per capita methods	Use existing study data
Acceptance by stakeholders	10	8	6	7	5	2	5	2
Explicit treatment of uncertainty	5	7	5	5	7	2	2	2
Underpinned by valid data	7	6	4	5	3	2	2	2
Transparency and clarity	7	7	6	5	2	2	5	2
Appropriate to level of risk	7	4	5	9	10	7	2	2
Logical and theoretical approach	7	8	5	7	5	2	5	2
Empirical validation	5	8	5	7	5	2	5	2
Explicit treatment of factors that explain HH consumption	8	8	5	8	5	2	2	2
Flexibility to cope with new scenarios	5	7	5	7	5	2	7	2
Weighted score		**429**	**315**	**411**	**315**	**157**	**234**	**122**
Ranked		**1**	**3**	**2**	**3**	**6**	**5**	**7**

The analysis of the RAG matrices showed that regression models scored highly across high, medium and low vulnerability water resource zones. Major consumption group (MCG) modelling generally came second, although for the low vulnerability zones, this was a marginal second place and MCG modelling could potentially be used in these zones. However, there is a benefit in terms of consistency and efficiency of analysis if a single, coherent framework for forecasting consumption across the six Thames Water WRZs is achievable.

The ability to use standard statistical modelling techniques and calculate model errors, which is offered by regression modelling, is particularly desirable in the higher risk zones in order to place confidence limits on consumption forecasts and quantify risk.

Micro-component forecasts are required for the WRMP19 tables and can be split out from the final forecast if it is Multiple Linear Regression (MLR) based. A fundamental principle of the MLR development is spatial validation; applying the model to all WRZs improves the performance of the model. For this reason, as well as work efficiency, MLR modelling, where used, is the preferred method in all zones.

Multiple linear models were selected as the most appropriate consumption modelling approach for forecasting consumption by 25 years, and extrapolating this to 2100 in all of Thames Water's water resource zones.

3 Multiple linear regression

3.1 Background

Consumption forecasts are commonly provided on per capita consumption (PCC) and per household consumption (PHC) basis. Linear modelling can use either metric.

In the case of PHC modelling, occupancy becomes an explanatory variable, and PHC is composed of a consumption allotted to the house on the basis of its characteristics, and an additional consumption assigned to each occupant. PCC modelling assigns a different consumption value per person on the basis of the characteristics of the property they inhabit. In the former case, the model is property driven, which aligns with the data collection based on household meter reads.

The latter case introduces the error associated with the household occupancy figure into the model at the very first step. If the model is based on PCC, the PCC is calculated from estimated occupancy (for which there is an error), so there is no part of the consumption modelling that is independent of occupancy error. This is a fundamental reason why linear modelling by PHC outperforms PCC. In forecasting, the same applies: all the error in population forecasting is propagated through the zonal forecast if it is based on PCC.

Modelling by PHC makes occupancy-driven household consumption components implicit in the model whereas PCC-driven modelling would need to incorporate a correction for changing occupancy rates in PCC forecasting.

A further reason for PHC modelling is that property numbers and types in the model trainer data are better understood than population, and are not dependent on questionnaire data.

For these reasons PHC is used as the basis for aggregating up to a zonal consumption modelling.

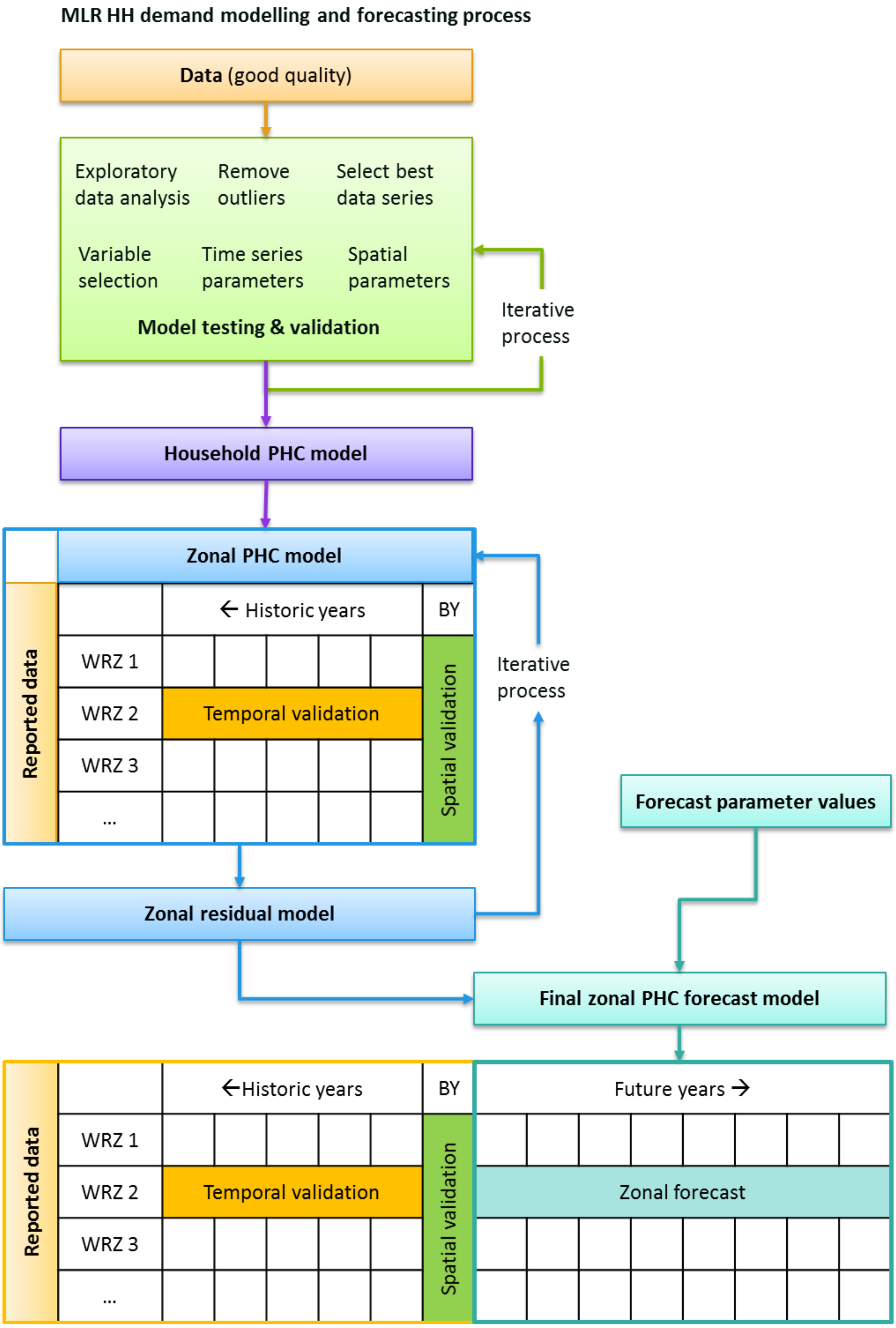
3.2 Approach to multiple linear regression modelling

Following the review of WRZ ‘problem characterisation’ and data availability, multiple linear regression (MLR) methods were identified as the preferred modelling approach to forecast demand in each of the WRZs.

Household consumption is affected by a complex mix of interacting drivers including: the make-up of the occupants (numbers, age, socio-demographics, their habits, practices and behaviours), the property type, whether they pay on a measured or rateable value bill, geography, etc.

The MLR approach uses standard statistical processes; these are applied in an iterative manner exploring the model errors and uncertainty at each stage until a useable and robust model is derived, if at all. Figure 2 gives a high level overview of this MLR process, which is explained in more detail later.

Figure 2 MLR Modelling process diagram



The resulting model has a number of model variables; each has a coefficient that is derived from the model and there is residual error term. The residual is essentially the consumption component that cannot be explained by the model variables. Residuals are used for estimating error and developing further modelling refinements.

One benefit of using MLR to create the household forecast, is the logical process in which the model is derived. The following process gives a more in depth view of the steps given in Figure 2.

1. **Obtain data** and explanatory variables from the Water Company.

2. **Select data** with which to build the model, based on;

   a. Sample size – sample needs to be sufficiently large so that extremes of the distribution can be modelled.

   b. Amount of historic data – long term availability of data is important for modelling trends and for data stability.

   c. Representativeness of population – does the sample adequately reflect the characteristics of the population it is trying to model?

   d. Number of explanatory variables – is there a sufficient amount of demographic data collected at household level?

   e. Age of data – Has the data been recently collected, and could external drivers have caused bias in the data since its collection?

3. **Exploratory data analysis** on the selected data set, to determine;

   a. The presence of outliers and how to deal with them;

   b. The distributions of the data, specifically the response variable;

   c. If missing values are present, and how to deal with them;

   d. The presence and removal of duplicate observations.

4. **Selection of variables** for inclusion within the model. Once the data has been analysed, and outliers and missing data are removed, both automatic and manual variable selection techniques (such as stepwise selection) are performed to identify variables which are significant in the model build.

5. **Identify variables which can be forecast,** and remove other variables from the model. It is likely that the 'ideal' model includes variables which cannot be forecast into the future, for example dishwasher usage. Therefore, a secondary version of the model is created which includes all significant parameters which can be forecast into the future.

6. **Test model assumptions** and validate the usage of MLR modelling. Using multiple linear regression requires that the data be tested for;

   a. A linear relationship between the response variable and the explanatory variable. This is verified by analysing a plot of the residuals vs. the fitted points.

   b. The expectation of the error term is zero for all observations, i.e. $\mathbb{E}(\varepsilon_i) = 0$ for all i.

c. Homoscedasticity – The variance of the error term is constant across the variables and over time. A plot of the standardised residuals verses the predicted values can show whether the points are equally distributed or not. If the variance is not constant, then the model uncertainty will vary for different observations leading to heteroscedasticity.

d. No multicollinearity, which assumes that the explanatory variables are not highly correlated with one another. Again, this can be determined using the standard residuals as well as looking at variance inflation factors.

7. **Model testing and validation** at household level, by way of coefficient resampling, and cross validation.

8. **Aggregate model to zonal level** so that zonal consumption figures can be derived as per the WRMP requirements.

9. **Zonal model validation,** similar to the household level validation, but using zonal reported figures. Again this is done using cross validation by excluding data by time period (years) and by zone to test the model spatially and temporally.

10. **Residual analysis** to determine if other factors which cannot be considered at household level (such as weather or climate effects) can be incorporated into a secondary model which will act upon the initial outputs.

11. **Trends and scenarios** are finally applied to the forecast based on the most likely scenarios for future behavioural and technological changes.

12. **Uncertainty** calculations are performed on the final forecast to give a 95% confidence interval for future predictions.

4 Review data availability

Now that the MLR methodology has been selected, this section reviews what data is available, how complete it is, and whether it can be used for modelling.

A range of potential data sources were identified at the start of the project, as described in Table 6.

Table 6 Potential data sources

UKWIR customer behaviour study	New data from smart meters
Market transformation programme	TW R&D studies
Domestic Water Use Survey (DWUS)	Smarter Home Visit data – skewed dataset. Water efficiency visits.
VMR & meter trial data	Wastage study
Peak factors study	Multi faith water use studies
Micro-component studies	

Discussions with Thames Water indicated that the most useful data for analysis were from the following sources, due to the number of demographics, sample size and amount of historic data, further explained in step 2 of Section 3.2 above.

- The Domestic Water Use Survey (DWUS);
  - a) Largest number of demographics
  - b) Representative sample (as confirmed by Thames Water)
  - c) Consumption data from multiple zones
  - d) Large sample size
  - e) 10 year timeline
  - f) Data fairly new at 6 months old.
- Smarter Home Visits;
  - a) Fairly large sample size
  - b) Few demographics
  - c) Average quality consumption values.
- Smart metering data.
  - a) No demographics
  - b) Useful for checking range of data (PHCs and leakage) against DWUS.

Of these, the DWUS data was likely to be most useful as it:

- Contained a representative sample of Thames Water household customers due to the way in which the sample was derived. This has been confirmed by Thames Water.
- Included a long-period of quality checked consumption data. Details of the automatic consumption checks can be found in Section 5;
- Information on individual properties; and
- Approximately annual survey data on their occupants, and the ownership and frequency of use of water using devices.

Therefore the DWUS data became the principal focus of our analysis for this project.

5 Exploratory data analysis

The principal purpose of the DWUS monitor is to enable Thames Water to estimate unmeasured household consumption. DWUS is a panel of customers who have voluntarily had meters installed but are charged on an unmeasured basis.

The DWUS data supplied by Thames Water from the 'DWUS' monitor contained daily flow records for just over three thousand properties over the period 2006 to 2016. A survey record associated with each day's summary flow data was supplied separately. The survey reference number ties the questionnaire data to the flow data for specific periods of time.

Table 7 gives an example of the types of demographic data supplied by Thames Water, with Table 8 focusing on variables specifically linked with property type. Property record data is, unlike survey data, assumed to be constant over the study period.

Additional variables were supplied, however it was not possible to discern exactly what the data had captured due to ambiguous file names. These variables were still included within the variable selection stage of the modelling procedure, and if significant, further details surrounding their meaning would be obtained. The total number of variables within each category is given in brackets in the table headings.

Table 7 Classification of survey parameters

Survey Variables		
Binary (27)	**Discrete (19)**	**Categorical (3)**
People at home in the day?	Car wash frequency	Occupiers income band (10 categories)
Own a dishwasher?	Dishwasher age	Occupiers hobbies (16 categories)
Own an economy dishwasher?	Number of adults	
Own a water fountain?	Number of sinks	
Cistern displacement device fitted?	Number of baths	
Own and use hosepipe?	Number of bidets	
Own and use irrigation system?	Number of cars	
Own and use jet washer?	Number of children	
Is the customer measured?	Number of dual flush toilets	

Survey Variables		
Binary (27)	**Discrete (19)**	**Categorical (3)**
Own an outside tap?	Number of occupants	
Home owner?	Number of power showers	
Own a paddling pool?	Total number of showers	
Have a pond?	Total number of sinks	
Have a sprinkler system?	Total number of toilets	
Own a storage tank?	Washing machine age	
Have a swimming pool?	Number of people in daytime	
Own a washing machine?	Number of people at weekend	
Own a water butt?	Number of standard showers	

Table 8 Demographics associated with of property details

Survey Variables		
Binary (1)	**Continuous (4)**	**Categorical (8)**
	Garden length	Age category (6 categories)
	Garden area	Property type (5 categories)
	Garden width	Postcode (2906 categories)
	Rateable value	ACORN group (17 categories)
		Resource zone (13 categories)

5.1 Data validity

Included within the survey data were validity flags indicating the period within which the survey may be considered valid. This is important, since the flow data is aggregated by year and by survey number, to ensure that different surveys and their results are classed as different data points for the household in question. Therefore, when calculating a households consumption, (PHC value), readings outside of the validity period were not considered, and resulted in the loss of 13% of the consumption data. The eliminated data fell across different time periods, which is shown in Appendix section 22.1, so can be considered to not have biased the results.

The eliminated data was also fairly constant over property type. However, there was a slight bias since flats had a higher invalidity rate of 16%, and houses 12% (within +-1% by property type subdivision).

However, the removal of this flow data only accounts for time periods for which the explanatory variables may not be valid, and does not account for outliers or erroneous data. Therefore, the next section analyses the flow data for inconsistencies and points to be removed from further analysis.

5.2 DWUS flow data

The DWUS data set has been maintained carefully over its lifetime and has been subject to extensive review and QA by Thames Water to ensure a ‘clean’ data set. Therefore, included in the summary daily flow data were automated daily flags for the metrics shown in Table 9. It was not necessary to remove all data flagged in this way, but allowed further examination of their prevalence within the data.

Table 9 Error flags provided by Thames Water

Flag Name	Description
Time Error	Incorrect/missing time stamp
Negative	Negative flow
High Flow	Flow over 1 l/s
Meter Check	Flagged if more than 10% difference between logged and measured flow
Repeat	Flag for multiple identical flow values, indicating a logger fault
Empty	Property empty
Leakage	Continuous flows over a pre-defined period (unknown)
Exclude from Analysis	Flag to exclude data Thames manual checks
Exclude Site	Flag to exclude data Thames manual checks

The first seven error categories in Table 9 have been combined with a single error code that can be disaggregated into individual error types (“MaxFlag”). This is important because leakage, for instance is of interest in itself.

There are 58 combinations of error flag, which when ranked drop off rapidly in importance in a log linear manner. Table 10 shows combinations ranking over 0.1%

Table 10 Top 12 ranked "MaxFlag" error categories

% of Total Sample	Sample average PHC l/day	Composite Error Flags	Difference from mean un-flagged consumption
80.53%	380.1	No error flag	0.00%
6.74%	359.5	Meter Check	-5.41%
4.32%	1123.0	Leakage	195.42%
2.75%	5.1	Meter Check, Empty	-98.66%
1.74%	659.3	Time error, Negative, High Flow, Meter Check, Repeat, Empty, Leakage	73.44%
1.09%	58.1	Time error, Meter Check, Empty	-84.73%
0.71%	15.7	Empty	-95.86%
0.54%	944.4	Time Error, High Flow, Meter Check, Repeat, Leakage	148.45%
0.46%	0.0	Time Error, High Flow, Repeat, Leakage	-100.00%
0.44%	1518.0	Meter Check, Leakage	299.35%
0.22%	112.8	Time Error, Empty	-70.33%
0.10%	41.8	Time Error	-89.00%

Meter check error flag arises where meter read departs by over 10% from the data logged consumption, and since it is one of the error codes that deviates least from non-flagged consumptions, it suggests that the error may lie with the manual meter read rather than the logged data in a large proportion of cases. We may assume that there is no systematic bias with meter read errors, as is also the case with faulty data loggers and flow lead faults that give rise to negative flows, repeats, and time errors.

There were two additional columns of exclusion flags supplied by Thames ('Exclude from Analysis' and 'Exclude Site') which included an element of inspection. These flags are designed to augment the automated 'MaxFlag', which takes care of the largest errors, as can be seen from the average consumption column in Table 11. This table is presented before any data is excluded using 'MaxFlag'. Therefore, a data point may pass QA using this method, but would subsequently be removed following a 'MaxFlag' error.

The PHC of data with both exclusion flags activated is actually very reasonable, unlike the PHC of data with neither exclusion flag activated where the PHC is extremely high. The two exclusion flags together account for 2.44% of data, of which 0.79% is also excluded by 'MaxFlag'.

Table 11 Exclusion flags

Exclude From Analysis Flag	Exclude Site Flag	Average PHC (l/prop/day)	Sample #	Proportion
FALSE	FALSE	13702	5564950	97.56%
TRUE	FALSE	1987	65261	1.14%
FALSE	TRUE	666	73879	1.30%
TRUE	TRUE	543	6772	0.12%

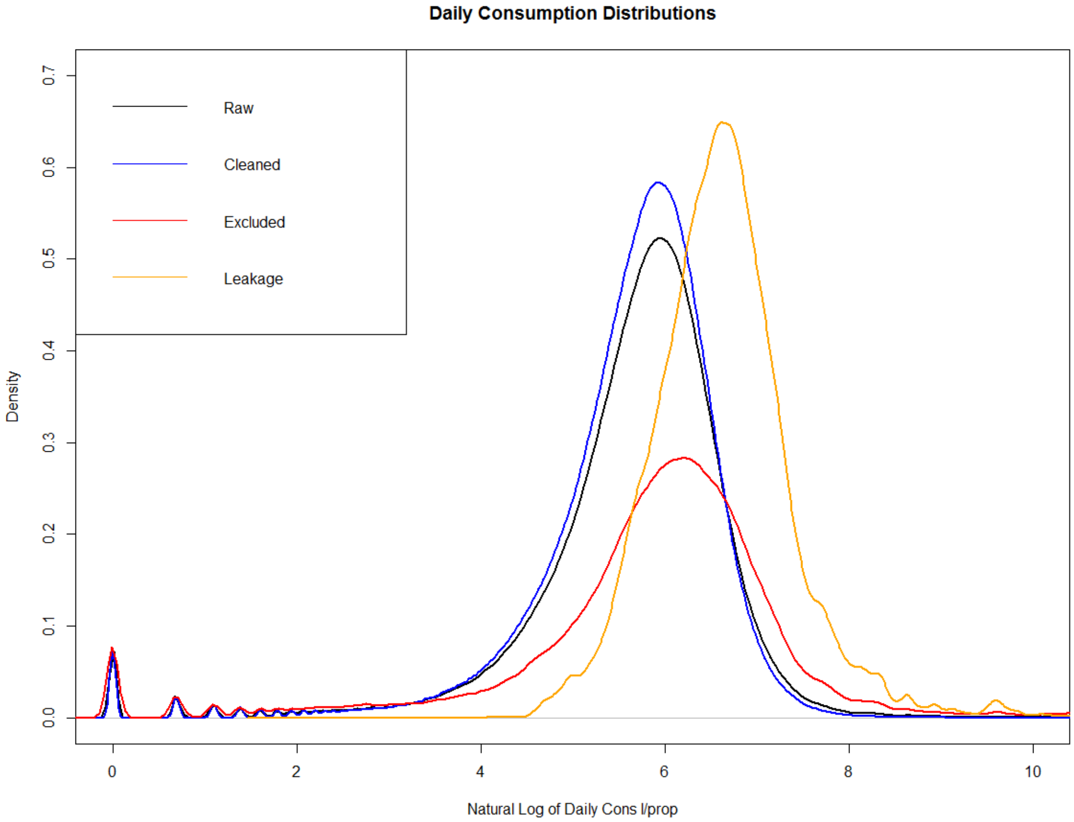
5.2.1 *MaxFlag removal*

It was decided that data with flags such as ‘Negative’, ‘Time Error’, ‘Repeat’ and ‘Exclude’, would automatically be removed since these figures clearly either contained impossible data (negative flows), flows as recorded by a faulty logger and therefore unreliable, or had the wrong time stamp applied and may result in incorrect calculation of PHCs. The other variables required more scrutiny.

The elimination of continuous flows (given by the ‘leakage’ flag) means that “legitimate consumption” without leakage is retained for analysis. This is important, since these flows may skew the results if leakage is biased toward households with, for instance a high per capita consumption (PCC).

Figure 3 shows the distributions of the daily flow before and after it is ‘cleaned’ by removing data with any flag from Table 9. This resulted in a total of 1,214,299 points being null or removed from the sample, corresponding to 21.26% of daily consumption data.

Figure 3 Distribution of daily flow data with and without leakage



Leakage influenced consumption is shown to be well described as a log normal distribution. The maximum single daily flow of any record in the cleaned data set is a figure that is large, but corresponds to a 24hr flow rate of 0.4 litres a second; a rate that is well within the capabilities of a 15mm domestic meter (over 1 l/sec).

The leakage flagged figure is double this, and the excluded data is impossible at 30 m^3 per second.

Given these observations, and the quality of the data available for error flagging, we do not consider that removal of further data points from the cleaned sample is necessary or desirable at this stage. Once the model has been built, we will use diagnostic plots to determine if any points are highly influential to the model, and further investigation of these points may result in additional data being excluded from the analysis.

Once the flow data had been checked, this was aggregated by year and survey number to create PHC readings per household. The flow data set then had the associated explanatory variables appended to produce a manageable data set. Since data had been removed according to the survey validity period, as well as the Thames Water error flags, Table 12 shows the number of properties by year that were included in the analysis.

Table 12 Proportion of data flagged invalid by year

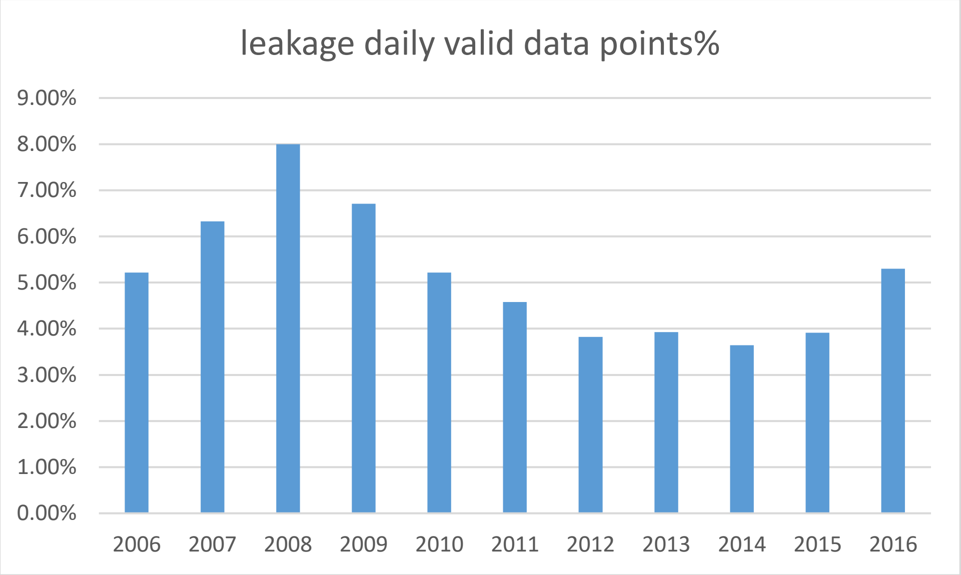
Year	# sites logged	# valid sites	Flagged Faulty
2006	2,407	1,589	34%
2007	3,569	2,524	29%
2008	2,626	2,082	21%
2009	2,335	2,001	14%
2010	2,269	1,922	15%
2011	2,242	1,959	13%
2012	1,975	1,670	15%
2013	3,910	3,124	20%
2014	2,828	2,322	18%
2015	2,245	1,740	22%
2016	1,327	809	39%
Grand Total	27,733	21,742	22%

Figure 6 shows that the standard error is much tighter for the middle years of the study, indicating fewer outliers for this period. The downward trend seen here possibly reflects a genuine downward trend in consumption that would be evident in aggregate error if those errors were symmetrical.

Larger standard errors at the beginning of the study period may reflect teething problems with the programme. Note that years 2006 and 2016 are not complete years, so are not commented on.

It might be expected that random data error should be proportionately constant year on year. Although a 'u-shaped' curve such as hinted at in Table 12 could be generated by the lifecycle of logging equipment, leakage is an external phenomena that we might expect to be random, although Figure 4 shows a peak in 2007 to 2009. This could be due to annual leakage variation, and there were a couple of hard winters in this period, or variation in the effectiveness of the automatic error flagging which may have a manual component.

Figure 4 Proportion of Leakage flagged valid data by year



Property empty is another parameter that shows an unexpected u-shaped curve (Figure 5).

Figure 5 Property Empty as a proportion of valid data

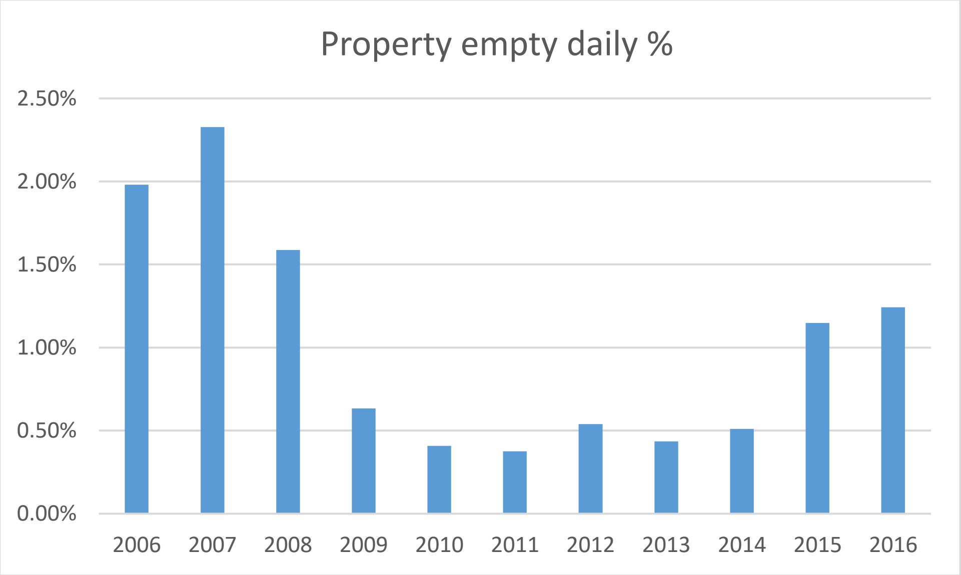
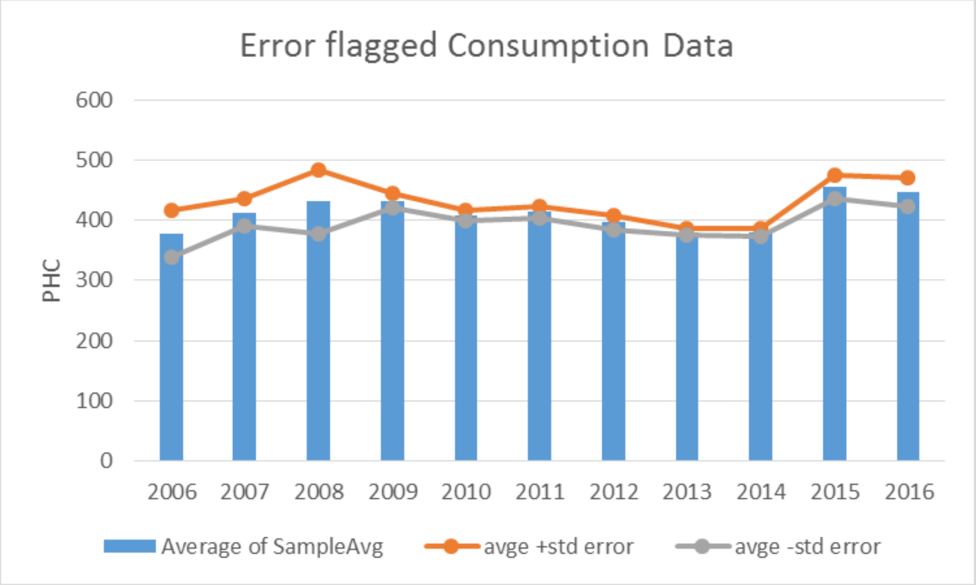


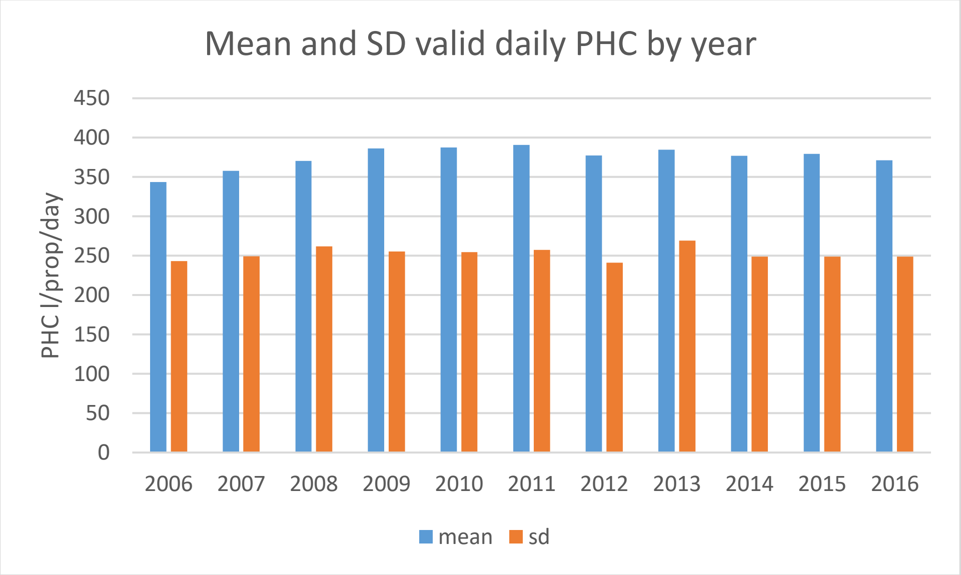
Figure 6 Time series of excluded data



In conclusion, it does seem that the scrutiny of flow data, automatic or manual, has been variable in quality over the study period although it is not possible to state what the best and worst years might be. Excluded data (shown in Figure 6) should intuitively show more variance when exclusion is more selective, however, the opposite applies.

This could mean that either that high variance errors are being missed in the middle years, or they were never present. Figure 7 indicates that variance of results in the accepted data is very uniform, which suggests that in general the flow data is of better quality in the middle years of the logging program.

Figure 7 Mean and standard deviation of valid daily data points by year



Modelling is predicated on observed and explanatory data, and although we can infer that the best quality flow data may occur in 2012 due to low variance in both excluded and accepted data, the real concern in data quality lies with explanatory variables, as we shall see in section 5.5

5.3 Duplicate data

In the base data as received from Thames Water, there were no duplicate site surveys, consumption records or site details. This was determined by conducting searches using functions within R.

In terms of the possibility of overlapping survey validity periods for example, all flow results were averaged by survey record and year, so consumption would not have been inflated for an individual property record if there were overlapping periods.

5.4 Missing data

Once the data had been formatted in a manageable database, initial Investigations indicated that some variables were incomplete due to missing data. Because “line deletion” was used to remove ‘faulty’ data, a whole record (i.e. the full row of a survey responses for the year in question) would be lost if a single variable was absent in the selected variables.

Therefore, initial modelling concentrated on determining the best mix of available variables to ensure a robust time series model with an emphasis on easily obtained data. If a selected significant variable has sparse data, then further investigations into gap filling would take place.

Missing data by year

In addition to looking at missing data within a variable, it is possible to consider the amount of missing data per year. The DWUS data set contained data from 2006 – 2016, each with differing numbers of household records.

There was more missing data in some years compared to others, therefore the initial modelling strategy was to pick a year with a lower amount of missing data, and to use this year to determine the significance of various variables before deciding whether they ought to be included in the model.

All variables were examined by year to establish their completeness. A few were rejected, and a number examined more closely, with examples of this given in the next section.

It became clear at this stage that due to the apparent use of default zeroes (particularly) where data had not been updated for a survey, analysis of missing survey data could not be undertaken without reference to data validity.

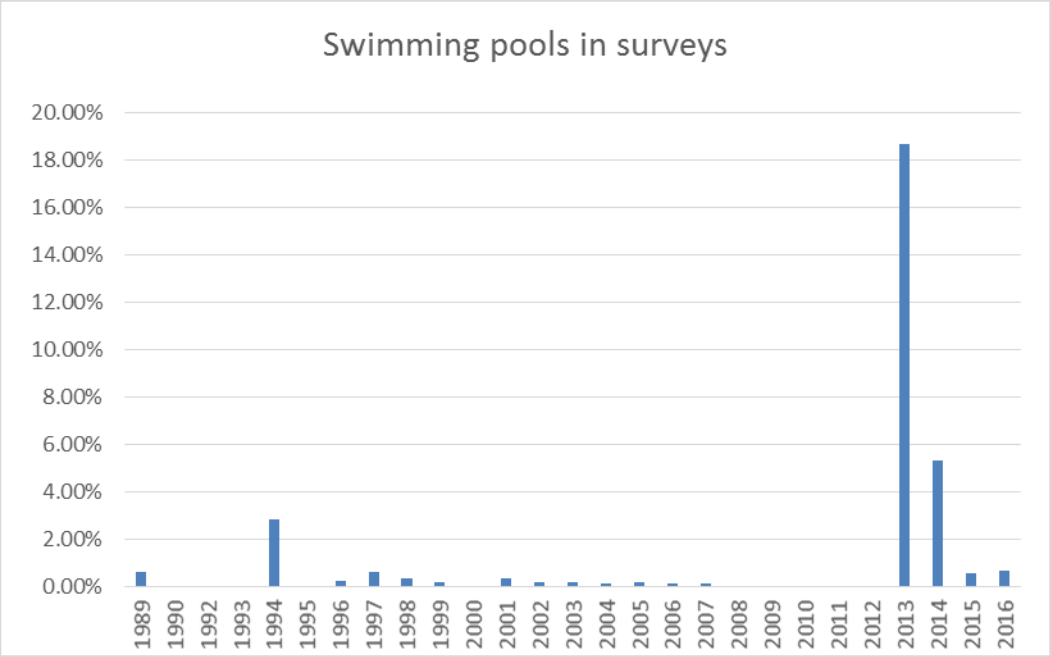
5.5 Survey Data QA

The survey data was thoroughly quality checked to determine which variables should be kept within the data set before variable selection took place. Different variables were rejected through different methods, with examples given below. Although, some variables could be rejected without close scrutiny, the majority of the data was closely inspected for errors.

Rejection without close examination

The first level of QA was done without close examination, which highlighted variables with very obvious faults. As an example, the variable 'swimming pool' was rejected as a viable parameter for two reasons. First, the erratic and unbelievable timeline proportions of ownership, shown in Figure 8, and second, the tiny sample size in other years.

Figure 8 Proportion of surveys citing swimming pool ownership



An ownership percentage of over 18% in 2013 but only 6% in 2014 seems unlikely. Therefore this variable was removed from the data set.

Rejection after closer examination

Most other variables required closer examination before a determination could be made for whether the variable should remain.

To illustrate this further, Figure 9 and Figure 10 show the change in proportionate home ownership between 2007 and 2013. This demonstrates that this variable is not credible, given that with a total of 3,633 properties having ever featured in the study, the majority of current study properties would have been subject to surveys. Since there is no data to support any of these years being accurate, and the parameter would perform inconsistently in time series, the parameter of home ownership was rejected after consideration.

Figure 9 Mean home ownership proportion from survey data

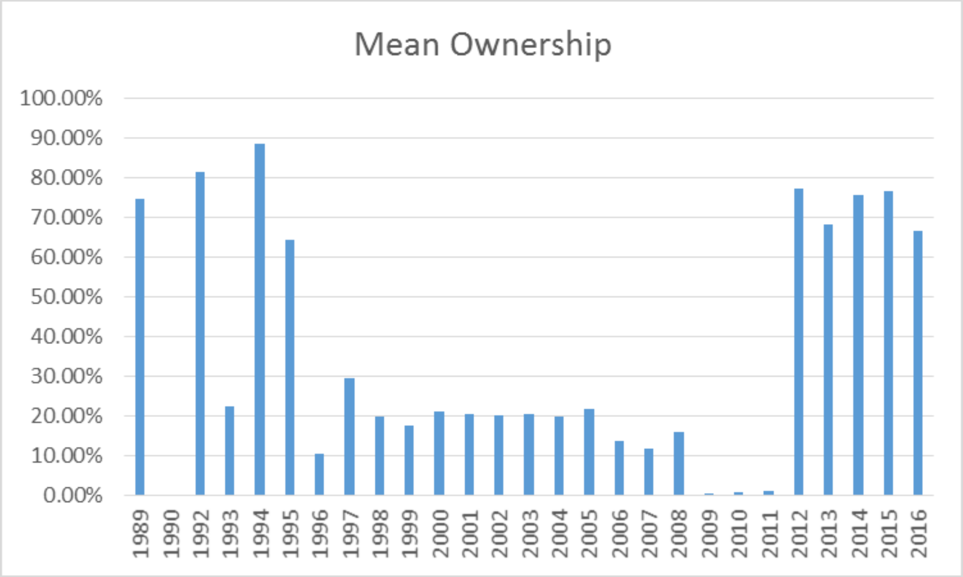
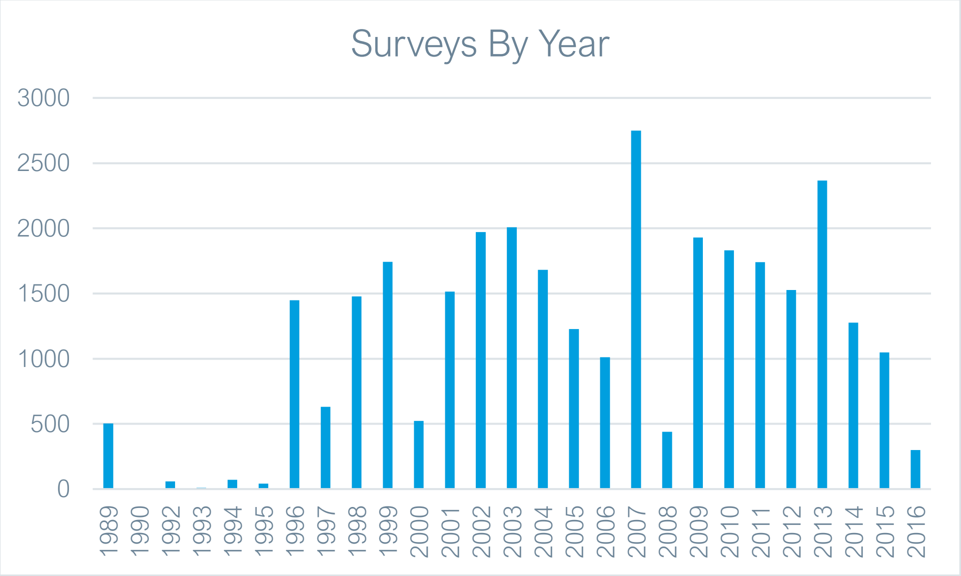


Figure 10 Number of surveys by year



Examination of variables that had shown significance in exploratory modelling were sometimes considered useable only for certain years. This both undermined confidence in the variable and precluded its use from time series modelling within the DWUS data set. (Figure 9)

Table 13 below shows the QA conclusions following the analysis of variables, only showing the variables which presented enough significance in the exploratory modelling to be considered for model inclusion. One potentially strong binary variable that had to be rejected was home ownership, as well as erratic ownership by year, as summarised in Figure 9. Visual inspection of this variable showed blocks of default values infilled as well as unlikely transitions between statuses in blocks arranged by survey number.

Table 13 Conclusions of parameter analysis

requires more analysis	Conclusion	Use unaltered for 2006/7 parameter significance?
washing machine	data up to 2007 good except for block 40 default zeros in 2007. 2009, 2013 both good	Y
dishwash	data up to 2007 good except for block 40 default zeros in 2007. 2009, 2013 both good	Y
dishwash age	Age not updated, so annual update from any decrease is necessary. Better to convert first appearance of age in time series to year of manufacture which remains constant until drop in age. Data not collected from 2008. Age Update caveat must also apply to washing machine age	N
storage tank	data up to 2007 good except for large blocks of default zeros in 2006, 2007; 2013 OK.	N
outside tap	data up to 2009 good. Use 2013 only if no other data available	Y
paddling pool	data up to 2007 good except for large blocks of default zeros in 2006, 2007; 2013 OK.	N
swimming pool	data up to 2012 good. 2013 onward data spurious	Y
hippo	data up to 2007 good. for gap filling use 2013, then 2009 as last resort	Y
water butt	data up to 2005 appears good. 2011 to 2013 appears good.Rest of the data afflicted with preponderance of default zero	N
irrigation	data up to 2012 good. 2013 onward data spurious	Y
trigger nozzle	data up to 2007 good except for large blocks of default zeros in 2006, 2007; 2013 OK.	N
at home in day	Use 2006 data only. Block of about 150 default zeros included	(Y)
#student away term	data up to 2007 good except for large blocks of default zeros in 2006, 2007; 2013 OK, 2009 OK ish. Cant really extrapolate this metric across years with good success rate	N

5.6 Correlations/covariance between variables

Due to the very large number of available parameters, which contained both binary, numeric and categorical data, formal analysis of covariance between model variables was put off until unimportant variables dropped out of the analysis.

Informal insights into covariance were gleaned from the impact on coefficients with the stepped removal and addition of other variables.

6 Model build strategy

The objective was initially to build the best model possible to describe the DWUS dataset using any available variables. This would provide a benchmark for consumption modelling performance using the best data set available.

The availability of variables across the Thames area would then be considered and necessary compromises to the model would be made to build a model that could be extrapolated to WRZ level. In this way an objective assessment of model performance at WRZ level could be inferred.

The performance of the model with the addition of new variables would be assessed using data from the year with the highest quality data (2007); the most recent year with high quality data (2013); and all intervening years taken together. Considerations of data availability at zonal level would then likely force compromises in model build.

Once the variables of interest were established, separate models for the Thames Valley and London would be built, given the high level of concern in the London WRZ.

When analysing the available data, it was noted that the measured set of properties in the DWUS (see Table 14) comprised of a 'self-selecting' group, in the sense that these people chose metering with a much larger prior knowledge than the standard optant or progressive metered households. For this reason, these properties were excluded from the analysis and an unmetered model was derived to model consumption.

Building a model using unmeasured consumption enables the behaviour of the unmeasured properties to be built into the model. The unmetered population of today is the metered population of tomorrow, so understanding the drivers of consumption in this cohort is a useful and necessary task for predicting metered consumption. As we move through time, metered homes will become the dominant cohort, so future forecasts will need to gather and analyse data from the progressive metered households to understand the drivers of consumption in these new cohorts.

Once metered, behavioural tendencies tend not to alter very rapidly, with the majority of savings stemming from reduced losses/leakage. Therefore, introducing a simple coefficient to scale the consumption values which have been modelled from the unmeasured behaviour, to a level consistent with the reported metered consumptions is a logical approach.

The 'self-selecting' metered group is therefore unsuitable to build an independent model, but is perfectly sufficient for understanding the potential savings, or scale difference, compared with the unmetered group.

Therefore, measured properties would be introduced as an adjustment to the unmeasured model, if the difference was shown to be significant. Details of this analysis are shown in Section 13.1.

Table 14 Annual data points available for model build

Valid data points	Measured	Unmeasured	Total
2006	375	1,214	1,589
2007	333	2,144	2,477
2008	79	2,002	2,081

Valid data points	Measured	Unmeasured	Total
2009	68	1,933	2,001
2010	49	1873	1,922
2011	87	1,872	1,959
2012	33	1,637	1,670
2013	30	3,094	3,124
2014	33	2,289	2,322
2015	28	1,712	1,740
2016	19	790	809

7 Variable selection strategies

Once the initial exploratory analysis has been completed, and the initial data set formatted, the variables for inclusion within the model are determined. Due to the large number of demographic variables supplied by Thames Water in the DWUS data set, a combination of automatic and manual selection techniques were used.

Selection of these variables is hierarchical, with the most important used first. Where variables conflict (i.e., are used for describing the same thing) the stronger variable (the one resulting in a higher adjusted R-squared model), is preferred.

The objective of this project is to produce a time series model that is able to forecast annual consumption. Therefore, the individual survey data were split into individual years to produce a time series baseline for trend analysis. Flow summary data, survey and property details were imported into the 'R' statistical software package.

Initial modelling, based on consumption, property and survey data from 2007 was undertaken in three principal ways using both data from all complete years (2007 to 2015), as well as 2007 and 2013 in isolation as being the years with definitive survey results. Although not all variables were considered viable for both 2007 and 2013, they were nonetheless examined for significance.

The methods used for variable selection are as follows:

- Unconstrained Optimisation: Using automatic selection based on libraries in R, this allows the optimising engine to select from the full range of potential variables after known problem variables were removed, in Section 5.5.

- Basket Optimisation: This uses the same automatic method, but uses a parameter basket approach which groups together similar variables to produce a two stage constrained optimisation.

- Manual optimisation: Carried out using the insights gained from the automated optimisation processes, ANOVA analysis summaries of various combinations of parameters, and covariance and correlations studied in the exploratory analysis stage.

Multiple methodologies were used, baselining significance of variables against stripped back models. Baselining variables included occupancy as well as more complex models.

The option of modelling PCC instead of PHC was explored, but dropped because of an initial consistent lack of success. The reasons for this are set out in section 3.1.

7.1 Unconstrained variable selection

Forward and backward selection, as well as stepwise selection was used within the R library 'leaps', which introduces variables one by one with the aim of minimising the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC) with considerations taken for model size. BIC tends to penalise larger models more heavily, and so the number of parameters influences the model selection.

Variable selection allows the optimal model to be determined, however not all selected variables will be useful in the final model. An example of this is the presence of an outside tap. Unlike property numbers, metering or population, we do not have a forecast of the number of outdoor taps in the future, or a good idea of how many are in each zone by measured and unmeasured property type. Therefore this parameter would not be appropriate to include in the final model.

In the same way, causative rather than proxy parameters are preferred as the behaviour of proxies is more unpredictable, and may shift in ways that are not well understood both spatially and in forecast. As an example, dishwasher ownership is a strong parameter in predicting higher than mean consumption. Because the associated coefficient is larger than the consumption impact of owning a machine, it is probably a proxy for lifestyle. As dishwasher ownership becomes more commonplace it is likely that it will represent a wider range of lifestyles.

Similarly, the automatic variable selection methods may select variables with covariance, or are not independent. Although there is broad agreement between manual and automated parameter selection, the automated selection has no knowledge about underlying issues such as parameter stability, or the likelihood of obtaining a forecast, so the final cut must always be through manual selection.

7.2 Unconstrained step optimisation

As previously explained, posting the full set of available parameters into the forward-back step optimiser for parameter selection can be ineffective, since parameter selections can be made for proxies, dependent variables or parameters which cannot be forecast. The model can also be ineffective when applied to a different sample to that from which it was derived. The reason for this threefold.

1. Categorical variables that carry lifestyle information must be left out because of the geometric increase in optimisation complexity introduced by multiple categorical variables.

2. There is a high degree of covariance in these variables, and data gaps disguise the causative variables and produce an over-fitted model.

3. Representative sample size varies.

Table 15 presents the T values from the unconstrained step optimisation and R-squared values obtained through model iterations from the initial step selection. The second column shows parameter T values for the automated parameter selection and is the ratio of a coefficient to its standard error. The initial selection of parameters was made using 2007 to 2015 data, parameters were then removed on the basis of their irrelevance to the model. It can be seen in Table 15 that individual parameters become stronger as the number of variables reduces through manual selection through iterations 2 to 4. The strongest parameters are number of adults and number of children, followed by IBP flag and rateable value.

Removing three parameters in the 3rd iteration (highlighted in yellow) does not noticeably affect the R-squared value of the model (0.41 compared with 0.40). This illustrates that the initial parameter selection results in an over complex model and is prone to overfitting.

The fourth iteration model was then applied to 2007 and 2013 data only, exposing the weakness in time series of some of the selected parameters. Given the renewal of survey results in 2013 it is interesting to note that the R-squared value for the 2007 model was considerably better. This indicated that testing marginal model improvements using the 2007 data set could be advantageous.

Table 15 Parameter T-values and model iterations

Average of T value	by iteration				2007 Data Only	2013 Data Only
Parameter	1st selection by step	2nd iteration	3rd iteration	4th iteration	4th iteration	4th iteration
(Intercept)	0.082	0.828	1.196	0.983	-1.794	0.242
NUM_ADULTS	38.167	42.507	44.714	44.702	24.286	10.348
NUM_CHILDREN	28.265	30.015	32.096	32.14	16.501	8.187
NUM_BATHS	4.628	4.601	4.532	4.956	3.521	1.985
TOT_SHOWERS	0.174					
NUM_PSHOWERS	-1.151	-0.938				
NUM_BASINS	-1.094					
NUM_DUAL_FLUSH	0.122					
NUM_TOILETS_1993	1.615	0.934				
NUM_WASTE_DISP	-0.11					
WASH_MACH.TRUE	-1.194	-1.735	-0.965			
STORAGE_TANK.TRUE	-0.717					
HOSEPIPE.TRUE	-3.372	-1.907	-1.161			
SPRINKLER.TRUE	-1.426	-1.548	-1.633	-1.695	-0.551	0.067
POND.TRUE	4.669	4.386	3.42	3.122	1.495	2.839
FOUNTAIN.TRUE	-1.448	-2.286	-2.715	-2.674	-2.422	-1.13
HALF_LOAD_WASH.TRUE	-1.59	-1.312	-1.183			
SOFTENER.TRUE	3.253	3.644	3.853	3.802	2.219	0.602
DISHWASHER_ECONOMYTRUE	5.648	4.571	4.42	4.233	2.619	0.764
WATERING_CAN.TRUE	-1.924	-2.434	-2.565	-3.027	-1.146	-1.514
INCOMEBAND	-1.52	-1.52	-3.063	-3.512	-2.308	-0.986
RATEABLE_VALUE	8.046	8.225	9.48	9.899	4.008	3.108
IBP Flag	17.928	19.22	19.098	19.139	9.718	5.937
Rsq	0.4161	0.4124	0.404	0.4038	0.424	0.334
Missing	17663	16686	15850	15850	636	2902
2007 data only Rsq	0.448	0.428	0.423	0.424		
2013 data only Rsq	0.35	0.337	0.334	0.334		

As a result of the automatic variable selection method and the unconstrained methodologies, the model variables which appeared in all optimisations are given in Table 16, and will be referred to as the 'base model'. This was used to determine the effectiveness of various categorical socio-demographic parameters.

Table 16 Base model output, for use to test the inclusion of other variables within the model

$$HH\ consumption = 14 + (108 \times no.adults) + (76 \times no.children) + (0.3 \times RV) + (195 \times IBP\ flag)$$

Residuals:	**Minimum** **-1552.26**	**1Q** **-105.06**	**Median** **-24.73**	**3Q** **74.52**	**Maximum** **1527.13**
	Coefficients:				
Parameter	**l/prop/day**	**Standard**	**t value**	**Pr(>|t|)**	**rating**
Intercept	13.66	6.07	2.25	0.0244	*
Number of	108.05	1.89	57.19	<2e-16	***
Number of	76.09	1.87	40.66	<2e-16	***
Rateable Value	0.3	0.02	14.59	<2e-16	***
IBP Flag	194.94	6.97	27.98	<2e-16	***

Residual standard error: 188.4 on 11851 degrees of freedom

Residuals:	Minimum -1552.26	1Q -105.06	Median -24.73	3Q 74.52	Maximum 1527.13
	Coefficients:				
Parameter	l/prop/day	Standard	t value	Pr(>|t|)	rating
(10813 observations deleted due to missingness) Multiple R-squared: 0.3977, Adjusted R-squared: 0.3975 F-statistic: 1957 on 4 and 11851 DF, p-value: < 2.2e-16					

Categorical variables were subsequently introduced into this model individually, with their effect on the model R-squared shown in Table 17. Only ‘ACORN type’ and ‘Property type’ had a beneficial effect on the model by increasing the R-squared from 0.398 to 0.406 and 0.404, respectively. As the ACORN parameter has 17 categories, there was a lot of underrepresentation in certain groups, so ‘Property type’ with just five categories is preferred.

Another advantage of ‘Property type’ over ACORN clustering is that the former is literally “bricks and mortar”, and “concrete”. Socio demographic clustering is much more subjective, much more difficult to forecast and because it requires a lot more data, dependent on a much smaller sample than property type.

Table 17 Effect of categorical variables introduced to base model

Alternative categorical variables	Model R-squared
Property type	0.404
ACORN type	0.406
ACORN category	0.398
Old ACORN	0.372
Socio-economic status	0.398
No Category - base model	0.398

7.3 Categorical covariance

Assumptions about parameter independence and covariance were tested using chi-square tests, the outputs in Table 18 show that the chi-square statistic which should be less than the critical value to support the assumption. Parameter combinations with more degrees of freedom are expected to have higher chi-square statistic so the ratio of chi-square to critical value is given as a comparative measure across tests.

The worst ratios in the table are between property type against Acorn type ID, justifying the decision not to include both, and also property type against IBP flag, justifying disaggregation of property type by IBP flag.

No results prove the assumption of independence, although some are clearly better than others.

Table 18 Chi-square test results

Parameter 1	Parameter 2	Chi square statistic	critical value @ 95% confidence	Ratio Chi Statistic to critical value	Assumption tested
prop type	# adults	3489	60.5	58	no correlation
prop type	New acorn type id	10189	83	123	no correlation
prop type	# children	664	41.3	16	no correlation
# Adults	# children	1443	38.1	38	no correlation
prop type	car wash binary	195	3.8	51	no correlation
prop type	daytime occupancy	95	9.5	10	no correlation
prop type	Dishwasher ownership	968	9.5	102	no correlation
prop type	IBP flag	629	3.8	166	consumption is the same

7.4 Constrained step optimisation

In order to narrow the selection and address the issue of covariance, the variable selection was carried out in two stages. The automatic optimisation in the “leaps” R library was offered occupancy plus one of five baskets shown in Table 19, each containing similar explanatory variables (for example, the number of bidets and the income band of the owners may be reflective of a person’s lifestyle, and may be correlated). Table 20 shows the T values before and after the weaker components had been removed.

Table 19 Parameter optimisation baskets

Occupancy	Estate Size	Profligacy	Parsimony	Lifestyle
NUM_OCCUPANTS	NUM_BASINS	WASH_MACH.	WATERING_CAN.	SOFTENER.
NUM_CHILDREN	NUM_WASTE_DISP	OUTSIDE_TAP.	WATER_BUTT.	REVISED_ETHNICITY
NUM_ADULTS	NUM_CARS	FOUNTAIN.	HALF_LOAD_WASH.	NIGHT_SHIFT.
NO_PEOPLE_HOME_WEEKEND	TOT_SINKS	IRRIGATION.	DISHWASHER_ECONOMY	INCOMEBAND
NUM_STUD_HERE_TERM	TOT_TOILETS	NUM_PSHOWERS	HIPPO.	OWNER.
NO_PEOPLE_HOME_DAYTIME	NO_STD_SHOWERS	PADD_POOL.	TRIGGER_NOZZLE.	APPLIANCE_NIGHT_USE.
AT_HOME_IN_DAY.	NUM_TOILETS_1993	DISHWASH.	NUM_DUAL_FLUSH	NUM_BIDETS
NUM_STUD_AWAY_TERM	TOT_SHOWERS	SPRINKLER.		
NUM_BATHS		POND.		
		CAR_WASH_FREQ		
		JET_WASHER.		
		HOSEPIPE.		

The selected parameters from each basket were combined and entered into a second optimisation. The T values for the model output are shown in the first column of Table 20.

Table 20 Iterations 2-3 of constrained step optimisation model

	T values	
Parameter	2nd iteration	3rd iteration
People at home in the day?	-1.26	
Car wash frequency	0.91	2.09
Own dishwasher?	1.96	1.84
IBP Flag	8.14	10.80
Garden length	-0.36	
Hippo device fitted	-2.24	-2.48
Number of people home in the day	2.04	2.22
Number of people home at weekend	2.40	1.40
Number of adults	7.08	7.94
Number of sinks	-1.26	
Number of baths	1.18	
Number of dual flush toilers	-2.30	-2.80
Number of toilets pre-1993	2.06	2.38
Own a sprinkler?	-0.03	
Total number of showers	-0.74	
Total number of sinks	3.83	4.25
Total number of toilers	2.40	2.04
Use watering can?	-0.79	

The second constrained step iteration produced an over fitted model with irrational coefficients (e.g. negative coefficients for "number of basins", "total showers") and covariate parameters which outperformed the unconstrained step optimisation model.

Although the "parameter basket" iteration method produced a better model than the unconstrained step optimisation, (with an R-squared of nearly 0.6 for FY07 dataset), there are limitations mainly due to missing data. Once these issues are addressed, the core explanatory parameters identified are sufficiently stable to assess the significance of the available demographics.

7.5 Covariance of key parameters

Having established a shortlist of preferred parameters for time series modelling, the compatibility of these parameters was put under greater scrutiny and a final manual selection by step optimisation undertaken. The main concern was interactions between property type, IBP flag and other socio demographic variables such as ACORN.

The performance ceiling of this modelling approach is due in part to nonlinear relationships; particularly between the very important property IBP flag and property type variables. Additional exploratory analysis shows that this signal is disproportionately strong in category 3 (terrace) properties. These two parameters can be merged to form an alternative categorical parameter of property type/property IBP flag.

Table 15 showed that the gains from incorporating socio demographic categorical variables are marginal, and chi-squared analysis of property type versus ACORN type shows high covariance (Table 18). The same applies to IBP flag, but the fact IBP flag is both a more powerful predictor than ACORN as well as having just two categories means that the sensible approach was to combine the two categories into an IBP/property type variable.

Table 21 gives the difference between the IBP flagged and non-flagged properties segmented by household type. For example, 7.8% of the total detached homes in the sample are IBP flagged, meaning that 92.2% are non-IBP flagged. This shows that flats with an IBP property flag have a similar consumption to those without the flag (with PCC of 96% and PHC of 98% for flagged properties compared with non-IBP). However, in houses, particularly terraced, IBP flagged properties used disproportionately more water than un-flagged (1.94 times as much PHC and 1.52 times as much PCC). Therefore, the variables are covariant, and for this reason, the combination of property type and IBP flag was selected as a new variable.

Table 21 Property flag effect disaggregated by property type

Difference between IBP flagged and non-IBP flagged properties by household type	Detached	Semis	Terraced	Flat Block	Flat
Proportion of IBP flagged properties per household type	7.8%	3.9%	7.1%	8.5%	6.2%
IBP flagged occupancy as a percentage of non-IBP occupancy	125%	133%	128%	126%	126%
IBP flagged PHC as a percentage of non-IBP PHC	170%	157%	194%	121%	124%
IBP flagged PCC as a percentage of non-IBP PCC	136%	118%	152%	96%	98%

This step was the key point in creating the developed model. With the introduction of IBP-flagged-property-type as the categorical socio-demographic variable, a number of parameters in the third iteration constrained basket optimisation dropped out (see Section 7.2), and no alternative parameters were found that significantly improved the model.

This revised model is termed the “developed model” and considered to be the best model for time series DWUS data. Table 22 shows this in more detail.

Table 22 Developed model derived from all time steps

Developed Model	Min	1Q	Median	3Q	Max
Residuals:	-726.26	-101.13	-22.74	75.97	1376.7

Coefficients:	Litres/prop/day	Standard Error	T value	P-Value	Rating
(Intercept)	57.48	17.48	3.29	0.0010	**
Number of adults	101.01	3.23	31.27	0.0000	***
Number of children	78.01	3.38	23.09	0.0000	***
Non IBP Semi-detached	-43.48	10.48	-4.15	0.0000	***
Non IBP Terraced	-27.91	11.06	-2.52	0.0117	*
Non IBP Flat	-46.16	13.92	-3.32	0.0009	***
Non IBP Flat block	-13.10	12.66	-1.04	0.3009	
IBP Detached	141.26	27.78	5.09	0.0000	***
IBP Semi-detached	71.80	24.61	2.92	0.0035	**
IBP Terraced	292.01	20.81	14.03	0.0000	***
IBP Flat	-62.27	30.82	-2.02	0.0434	*
IBP Flat block	-59.89	34.31	-1.75	0.0810	.
Rateable value	0.28	0.04	6.97	0.0000	***
Number of people home in the day	14.44	3.19	4.53	0.0000	***
Hippo device fitted	-57.01	11.32	-5.03	0.0000	***
Car wash frequency	18.98	4.92	3.86	0.0001	***

Signif. Codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 172.7 on 3678 degrees of freedom

(18975 observations deleted due to missingness)

Multiple R-squared: 0.4493, Adjusted R-squared: 0.4471

F-statistic: 200.1 on 15 and 3678 DF, p-value: < 2.2e-16

Note: This model is able to exploit under 20% of data records

7.6 Base year selection

There are two base years to consider: the base year for calibrating model outputs, and the base year for building the model.

The model build at household level could use a number of years' data. It can be seen from the last two columns of Table 15 that the 2007 data produces a better model than 2013. The coefficients are also different, but we don't know to what extent these differences are error and to what extent they are genuine time series changes. Section 5 shows the analysis that was undertaken to decide which year would be the best to build the initial model with. However, the final model was built on all year's data as the increased sample size for small categories such as "IBP flag detached homes" outweighed the lower confidence at household level. This is illustrated in more detail in section 8.1.4.

8 Additional exploratory analysis

Now that the initial models have been built (base model and developed model), and the main variables have been identified; (number of adults, children, property type, IBP flag and rateable value (RV)), additional analysis is undertaken to review the attributes of these variables, and how using them will affect the model.

This analysis will focus on:

- Looking at missing data, and if gap filling methods can be used;
- The creation of new variables by combining or merging correlated variables together.

8.1 Missing data and gap filling

Parameters can be assessed by number of missing data fields, which can be by survey data, daily consumption data point or by property for immutable variables. These measures don't necessarily correspond with the total impact on model data points, some properties and some surveys being associated with more data points than others. As a high level view however, the table below shows data that's missing altogether.

Table 23 Missing survey/year fields for various important parameters

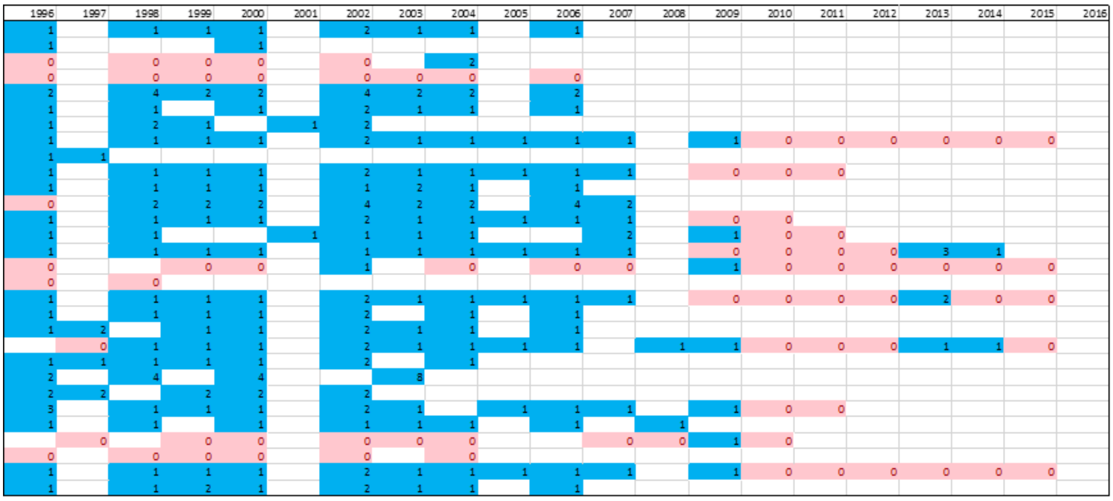
Parameter	Number of missing data points	Percent of total data set
Number of adults	44	0.14
Number of Children	153	0.49
IBP flag	0	0.00
Rateable Value	14,441	46.40
Property Type	14	0.04

Missing data is a simple way of looking at data quality. A more in depth look at categorical parameter data reveals data errors which are not evident from missing data analysis.

Figure 11 shows an example of errors in survey data. This particular example shows the sum of the reported number of showers by property survey by year. Each row represents a property, and each column shows different years. A blank indicates there is no survey record for the year, whereas a zero indicates that there has been at least one survey with no reported showers. Differences in yearly integer values (blue cells) may be due to multiple surveys and not due to errors as this analysis was designed to show the presence of zero bands (red cells).

It is evident from examination of these bandings for various variables, that there is a tendency to a default zero in years 2010 to 2012, and it would not be surprising if the model produced a low R-squared value when data from these years is incorporated.

Figure 11 Example of error bands



1996	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016
1		1	1	1		2	1	1		1										
1				1																
0		0	0	0		0		2												
0		0	0	0		0	0	0		0										
2		4	2	2		4	2	2		2										
1		1		1		2	1	1		1										
1		2	1		1	2														
1		1	1	1		2	1	1	1	1	1		1	0	0	0	0	0	0	
1	1																			
1		1	1	1		2	1	1	1	1	1		0	0	0					
1		1	1	1		1	2	1		1										
0		2	2	2		4	2	2		4	2									
1		1	1	1		2	1	1	1	1	1		0	0						
1		1			1	1	1	1			2		1	0	0					
1		1	1	1		1	1	1	1	1	1		0	0	0	0	3	1		
0			0	0		1		0		0	0		1	0	0	0	0	0	0	
0		0																		
1		1	1	1		2	1	1	1	1	1		0	0	0	0	2	0	0	
1		1	1	1		2		1		1										
1	2		1	1		2	1	1		1										
	0	1	1	1		2	1	1	1	1		1	1	0	0	0	1	1	0	
1	1	1	1	1		2		1												
2		4		4			8													
2	2		2	2		2														
3		1	1	1		2	1		1	1	1		1	0	0					
1		1		1		1	1	1		1		1								
	0		0	0		0	0	0			0	0	1	0						
0		0	0	0		0		0												
1		1	1	1		2	1	1	1	1	1		1	0	0	0	0	0	0	
1		1	2	1		2	1	1		1										

Section 7 identified the key explanatory parameters that appear to influence household consumption, based on DWUS data. This insight provided a focus on the variables to identify for gap filling. There are a range of options for gap filling data which include:

- Interpolation - Assumes that the data either side of a record describe the data in the centre. In this case that if there was a shower in 2007 and in 2009, then there was one in 2008.

- Imputation - Assumes that the missing data adheres to a rule that can be determined. Using the previous example, a rule for showers would be that once they are installed, they are not removed.

- Statistically neutral infilling - Assumes that missing data has the same distribution as the data present. For instance, rateable value of flats without a record have the same mean as flats that do have a record of rateable value.

The approach to gap filling will depend on the variable in question, as what is appropriate for showers for example may not be the correct approach for other variables such as cistern device "hippo", students' home term, or income band.

8.1.1 IBP flag

This variable has shown to be important in all models built so far, however this flag has not been successfully assigned to the entire population.

Properties which initially had an 'unknown' IBP flag (representing 3.3% of all properties) have been omitted from the base model to reduce the number of property type categorical variables to two.

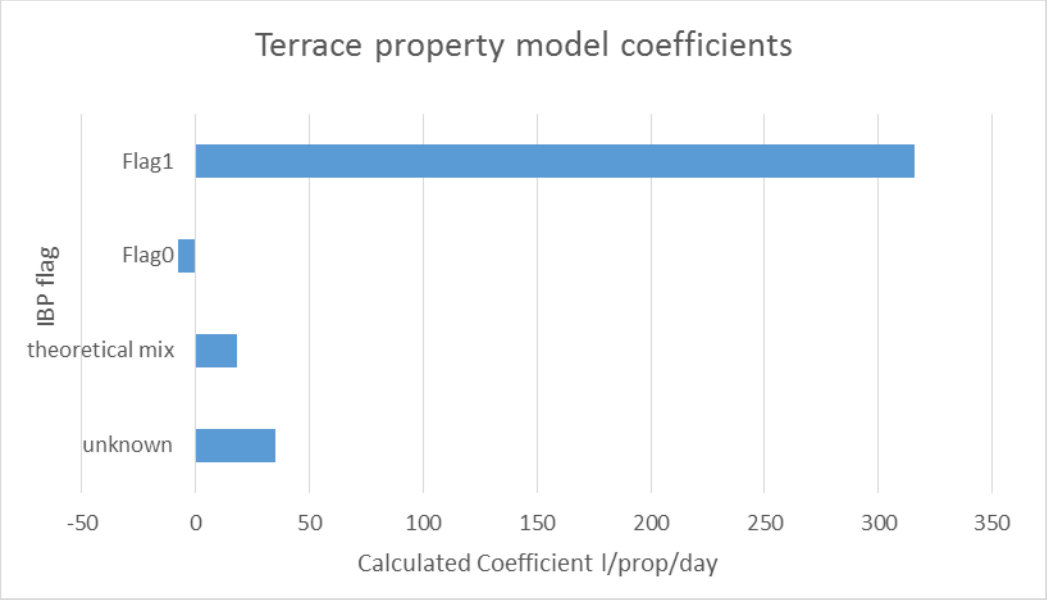
In testing the IBP flag 'unknown', the thesis was that 'unknown' is random and will represent a proportionate mix of IBP flag and non IBP flag categories as they occur in the DWUS population.

The category was included for a model run to examine the assigned coefficient and compare it with the expected coefficient if the random assumption is correct.

The expected coefficient was calculated from the flag 1 and flag 0 coefficients - i.e., that it would be weighted proportionately between them by the relative frequencies of flag 1 and 0 occurrences. This coefficient is termed "Theoretical Mix".

The largest IBP flag 1 category with the most significant sample size was terraces, and the results show that model calculated coefficient is a reasonable match to the theoretical mix coefficient.

Figure 12 Comparison of IBP Flagged vs Flag 0 and "Unknown" Coefficients



There is a slight observable increase in model quality when the unknown category is excluded, rather than put in with the un-flagged results. The R-squared value for the base model shows an improvement of 2.57% to 0.436.

8.1.2 Occupancy

The number of adults and number of children are the benchmark variables for both data completeness and data importance. As these variables are nearly complete, there is little point attempting to infer missing values for these parameters, especially as adult occupancy is the dominant variable and incorrect inference could alter the behaviour of sensitive peripheral variables. Therefore, the occupancy related variables will be kept as they are, for the purpose of the model build.

8.1.3 Property type

The data in the property type category, is effectively complete, with just seven records missing, as illustrated in Table 24.

Table 24 Property type data values

Property Type	Count	Description
1	594	Detached
2	1064	Semi
3	1123	Terrace
4	437	Flat in large block (>=6 flats in building)
5	408	Flat in small block (<6 flats in building)
Missing	7	
Grand Total	3633	

Therefore, the same as with the occupancy variables, no gap filling is to be performed on this data.

8.1.4 Rateable value

Unlike the other variables analysed, rateable value has 41% of its records missing. This does not translate directly to the number of missing data points in the base model, which is closer to 20%. This is because the analysis unit is property survey years; multiple surveys are assigned to a single property and properties with missing data tend to have fewer surveys assigned to them.

However, it is possible to infer a good estimate of rateable value from household type. Rateable value is reasonably well populated, and there are few variables which are better populated from which to select. Table 25 indicates that the standard error on the mean of rateable value inferred from property type is in the region of 1 to 2%.

Table 25 Relationship between rateable value and household type

Property Type	Count of Properties	Average of RV	Standard Deviation of RV	Standard error	% Standard error
Detached	594	404	138	5.66	1.40%
Semi	1064	295	85	2.60	0.88%
Terrace	1123	255	85	2.55	1.00%
Flat	437	258	85	4.06	1.57%
Block Flat	408	226	73	3.63	1.61%
Missing data	7				
Grand Total	3633	283	107	1.77	0.62%
Terrace, Flat combined	1560	256	85	2.16	0.84%

Running the base model with known and inferred values results in a reduction of the R-squared value from 43% to 35% with the imputed data. This is because the sample size is already very significant at over 50%. A sample size of over 99% does not make as much difference as incorporating gap filled data. Therefore, there seems no need to impute this data set.

8.1.5 Time series data

There are a number of time series variables used in the constrained model which may be imputed or inferred. Theoretically, these variables are updated when a new survey is carried out. Extensive scrutiny of time series records for various metrics revealed inconsistencies and banding that was clearly beyond random effects, and due to if/how specific questions were included in surveys.

For example, the number of people home in the daytime is a useful parameter which could be considerably populated within the model by data inference. These data have not generally been collected since 2009 and values in this period often include zeros instead of valid data. Gap filling would involve filling in values that are missing, and the use of a spurious zero default requires that we first remove false zeros by some method.

The technical challenge is that if there are genuine data in (say) 2011 and 2012, we can only recognise it as such because it is not zero. If we then delete all zeros and repopulate them with an imputed number, we are skewing the distribution. In this case, we must delete all entries to conserve the correct annual distribution.

Scrutiny of this data (Figure 11) suggests that the best form of gap filling for properties where some data exists would be to assume the maximum recorded value. With this approach, the chance of deviating from the maximum recorded value for any year is around 1%.

Where there is no data, so the highest historical value cannot be deployed, or the data originates post 2010, then a mean daytime occupancy could be used, or the mean daytime occupancy by some reasonable stratification, such as properties with the same total occupancy.

This example highlights the complexity and challenges of manual data imputation. This approach could also introduce unintended bias as a result of the judgement required to undertake this manual analysis.

These challenges were addressed by using an automated routine using the R library ‘mice’, which handles multiple imputation for time series data using an iterative process to preserve inter-parameter distributions. The only manual input required to the process was the deletion of banded erroneous data to allow infilling.

The base model and developed model were run with the imputation program, with manual infilling, and with line deletion.

Table 26 summarises the performance of the imputation programme (annotated as ‘Auto’ in the table) compared to manually imputed results. Auto imputation was able to generate relatively few additional data points. Based on these results it has been decided that imputing data is not worthwhile.

Table 26 **Performance summary of models**

Model	R-squared	Standard error	Missing Records	Methodology Employed
Model 1	0.425	183	11,212	Base Model
Model 2	0.336	206	76	Base Model rateable value imputed
Model 2 Auto	0.352	202	10,400	Base Model rateable value auto imputed
Model 3	0.448	171	20,418	Developed Model
Model 4	0.339	205	76	Developed Model imputed values
Model 4 Auto	0.356	202	19,956	Developed Model auto imputed values

8.2 Combining existing variables

During the development of the base model, it was found that there was a high level of covariance between the IBP flag variable, and property type. Section 7.5 describes the process of combining these two datasets into a single variable for use within the model.

Following this combination, and the derivation of the developed model, opportunities to combine other variables were explored, with the hope of further enhancing the base model. However, no other combinations were found to be effective.

8.3 Re-build models

Now that the gap filling options have been fully explored, the base year data examined and the new IBP/property type variable created the models are rebuilt for the final time.

Modelling consumption across the 2006 to 2016 period using the model derived from 2007 data produced a reduction in model fit from that achieved with 2007 data alone. The importance of some selected parameters in the initial modelling years was not necessarily carried through into subsequent years. This is no real loss, as the more complex models contained harder to obtain survey data and parameter selection is subject to more covariance.

Extensive analysis of the survey data demonstrated inconsistency in time series which is why model performance requires that model variables are dropped for the final model which is offset with the advantage that a much larger data set than the 2007 model alone. The survey data for occupancy is the parameter that seems to be updated most often, which is fortunate and not really surprising.

Stratification by property type, IBP flag and region (London/Thames Valley) exacerbates sample size to the extent that in isolation, 2007 data falls below a significant sample size for some property type coefficients, with the three compounded minority divisions being compromised (Table 27)

Table 27 Data points available for FY2007 only Thames Valley IBP flagged property types

Property type	Survey year sample size	property count	Data missing or excluding	Revised sample size for base model	Revised sample size for developed model
Terrace	15	15	2 missing RV 7 missing daytime occupancy	13	8
Block Flats	8	8	5 missing daytime occupancy	8	3
Detached	11	11	2 missing RV 6 missing daytime occupancy	9	5
Flat	5	3	2 missing daytime occupancy	3	1
Semi	14	14	2 missing RV 2 with leakage 9 missing daytime occupancy	10	5

It could be argued that if the 2007 model has a better R-squared then it is preferable, but it must be borne in mind that the model fit is partly dictated by the dominance of certain large sample categories, and if we wish to model areas with a higher concentration of low sample categories then the 2007 only model will perform less well compared to a model with a more significant sample of low sample categories.

Using the full range of available data also prevents exclusion of more recent data.

Extensive parameter optimisation and testing (as carried out for the 2007 data set) produced a more developed model which was robust across the study period, as summarised in Table 22. A large proportion of data was excluded by line deletion in this process. Section 8.1 describes the benefits of data gap filling techniques on model performance.

Incremental improvements in R-squared value are possible beyond the developed model, but the cost to robustness is unacceptable given the increasing instability of model coefficients and decreasing sample size when further parameters are included.

9 Forecast parameters

Since the final model will depend on the ability to forecast the chosen variables, forecasts were provided by Thames Water for the following data. Where forecasts were not available, but could be computed, a description of the algorithm (or method) is provided.

9.1.1 Household type categories

Thames Water provided data and forecasts of two household property characteristics. The first is a property type category, and the second was a binary classification of properties by IBP flag developed by Thames Water. These are two of the most powerful explanatory variables available, but correlations between them require that IBP flag must be used in conjunction with property type forming a new household type category (See section 7.5 Covariance of key parameters).

After the initial model build, and in light of the demonstrable importance for consumption modelling Thames Water supplied property type forecasts disaggregated by IBP flag.

9.1.2 Number of children

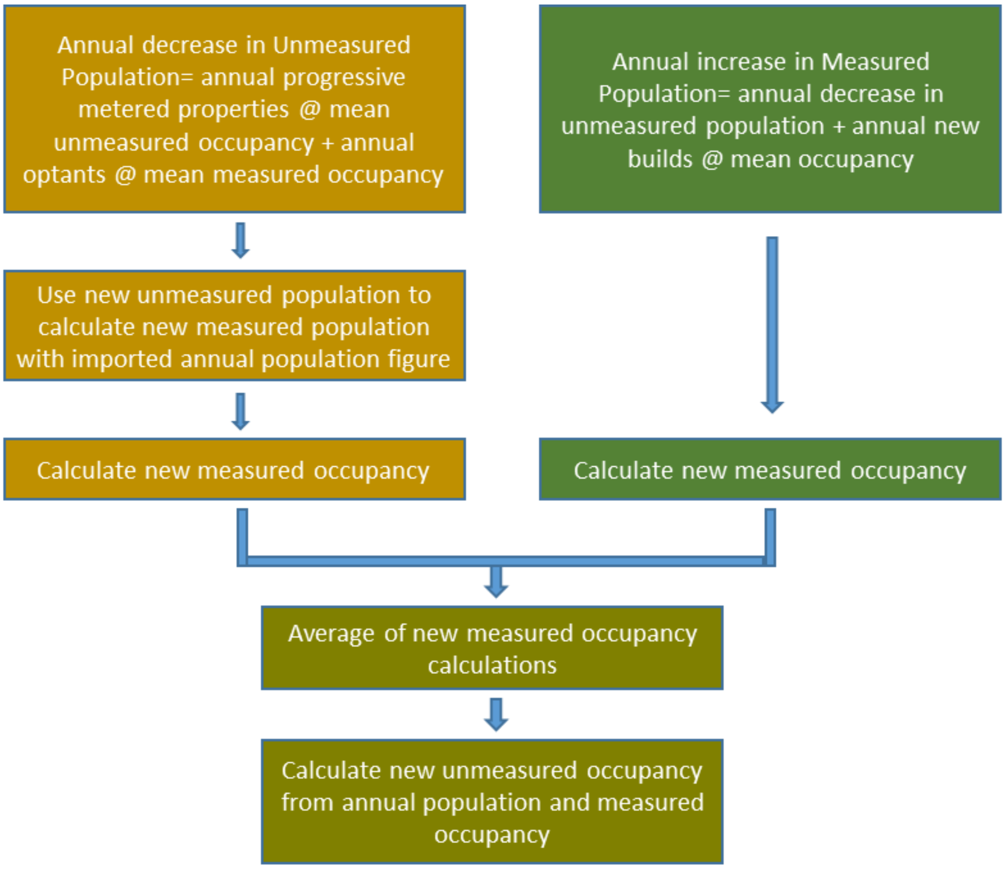
The variable for the number of children was not provided by Thames Water, so this figure was calculated by using ONS local authority level projections mapped on to a Local Authority to WRZ key provided by Thames Water.

9.1.3 Occupancy

Similarly, as forecasts were not available for both measured and unmeasured occupancy, Artesia have developed an algorithm to forecast the changing occupancy for measured and unmeasured population as meter penetration progresses. This algorithm is initiated by calculating the change in population expected from shifts in metering status and new build properties for both measured and unmeasured populations.

These are then used with forecast population changes to reconcile the two calculations against total population on an annual basis (Figure 13).

Figure 13 Occupancy Recalculation Algorithm



Annual decrease in Unmeasured Population= annual progressive metered properties @ mean unmeasured occupancy + annual optants @ mean measured occupancy

Use new unmeasured population to calculate new measured population with imported annual population figure

Calculate new measured occupancy

Annual increase in Measured Population= annual decrease in unmeasured population + annual new builds @ mean occupancy

Calculate new measured occupancy

Average of new measured occupancy calculations

Calculate new unmeasured occupancy from annual population and measured occupancy

10 Final model

Following a thorough analysis of model parameters, their interaction with other explanatory variables and their ability to be forecast, the final model parameters are given below. These parameters are the same for both the London and Thames Valley data sets.

$$Consumption = \alpha + \beta x_1 + \gamma x_2 + \boldsymbol{\delta} x_3 + \boldsymbol{\eta} x_4 + \upsilon x_5 + \varepsilon$$

Where:

x_1 Number of adults

x_2 Number of children

x_3 IBP flagged property type flag; either Semi-detached, terraced, flats, flat block or detached

x_4 Non-IBP flagged property type flag; either Semi-detached, terraced, flats or flat block

x_5 Rateable value (RV)

And the coefficients:

α Intercept

β Number of adults

γ Number of children

$\boldsymbol{\delta}$ Vector of coefficients for Asian property types; Semi-detached, terraced, flats, flat block and detached. The appropriate coefficient is used dependent on the value of x_3

$\boldsymbol{\eta}$ Vector of coefficients for Asian property types; Semi-detached, terraced, flats, flat block and detached. The appropriate coefficient is used dependent on the value of x_4

υ Rateable value (RV)

ε Error term

The values of the model coefficients are subject to change upon further identification of outliers in generating diagnostic plots in later sections. However, the process of selecting the significant variables is complete as the data has already been quality checked for sites including more obvious leakage. The coefficients of the final model are shown in Table 28.

Table 28 Final model output and coefficients

Residuals:	Minimum	1Q	Median	3Q	Maximum
	-1479	-104	-26	74	1440
Coefficients:	Litres/prop/day	Standard Error	T value	P-Value	Rating
(Intercept)	52.32	10.35	5.06	0.0000	***
Number of adults	102.73	1.92	53.41	0.0000	***
Number of children	73.68	1.85	39.90	0.0000	***
Non IBP Semi-detached	-21.62	6.72	-3.22	0.0013	**
Non-IBP Terraced	-11.23	6.83	-1.65	0.0999	.
Non-IBP Flat	-47.92	7.94	-6.04	0.0000	***
Non-IBP Flat block	-18.24	8.05	-2.27	0.0234	*
IBP Detached	262.77	18.69	14.06	0.0000	***
IBP Semi-detached	88.04	15.29	5.76	0.0000	***
IBP Terrace	316.89	11.83	26.78	0.0000	***
IBP Flat	-59.34	20.49	-2.90	0.0038	**
IBP Flat block	-0.43	19.60	-0.02	0.9825	
Rateable value	0.27	0.02	11.91	0.0000	***

Residual standard error: 184.4 on 11843 degrees of freedom

(10,813 observations deleted due to missingness)

Multiple R-squared: 0.4236,   Adjusted R-squared: 0.423

F-statistic: 725.3 on 12 and 11843 DF,   p-value: < 2.2e-16

11 Model assumptions

Now that the variable selection has been completed, the assumptions required to build a linear model are tested. Largely, the assumptions for building a regression model are the following:

- A linear relationship between the dependent variable and the explanatory variable. This is verified by analysing a plot of the residuals vs. the fitted points.

- The expectation of the error term is zero for all observations, i.e. $\mathbb{E}(\varepsilon_i) = 0$ for all i.

- Homoscedasticity – The variance of the error term is constant across the variables and over time. A plot of the standardised residuals versus the predicted values can show whether the points are equally distributed or not. If the variance is not constant, then the model uncertainty will vary for different observations leading to heteroscedasticity.

- No multicollinearity, which assumes that the independent variables are not highly correlated with one another. Again, this can be determined using the standard residuals as well as looking at variance inflation factors.

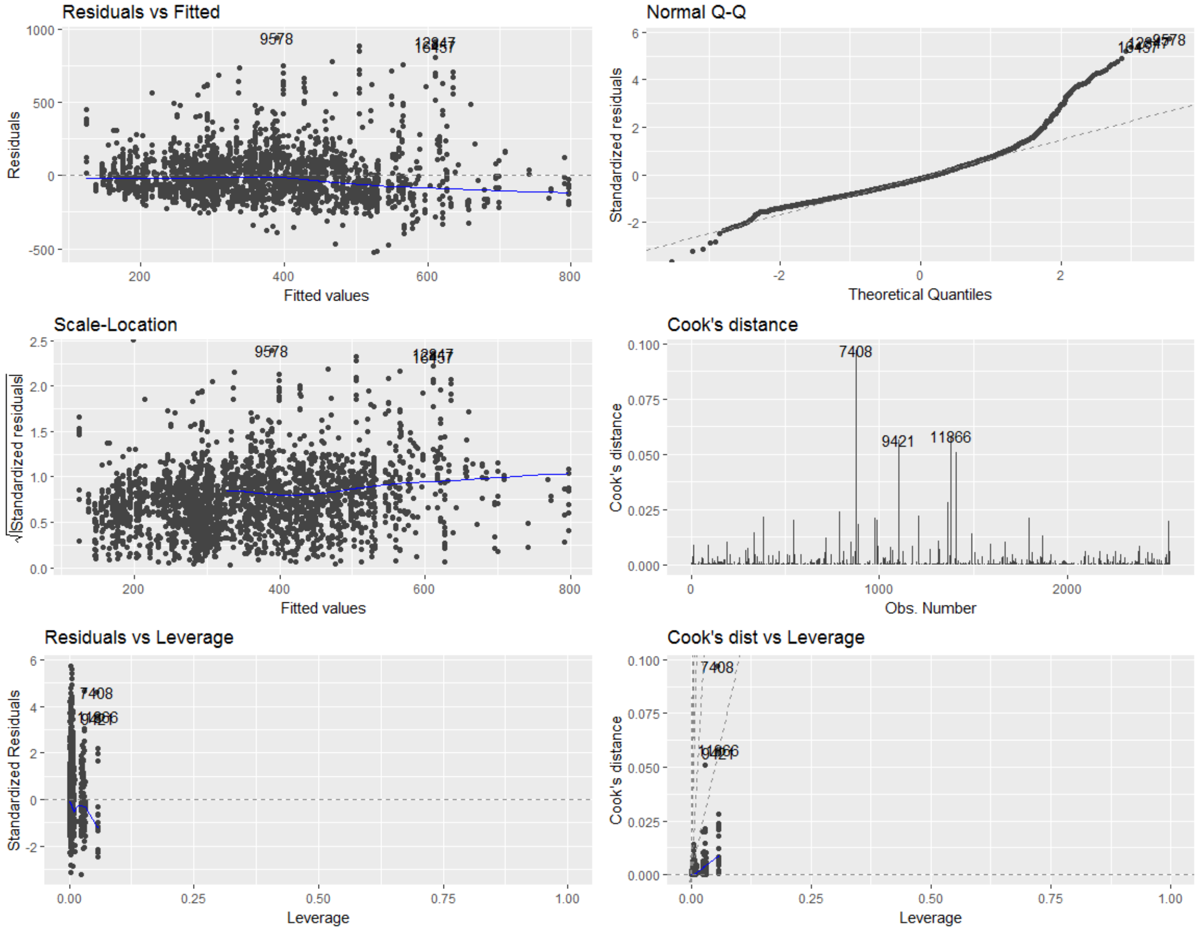
By satisfying these conditions, the Gauss Markov assumptions are met and the OLS estimators are best linear unbiased estimators. However in reality, real data rarely meets all of these requirements. This may mean that the estimator may not be best, but the choice of estimators with the data available is slim.

The reason that these tests are completed after the initial model selection is because of the number of variables available for inclusion within the model. It is not feasible or sensible to test for linearity amongst all variables, nor is it possible to look for correlations between every combination. Therefore, by completing the variable selection first, the number of diagnostic tests to be completed moving forward are markedly reduced.

To test the assumptions for the model, diagnostic plots were produced for both the London and Thames Valley models separately, to ensure that both are valid.

The results for the Thames Valley model are shown in Figure 14.

Figure 14 Model diagnostics for Thames Valley models



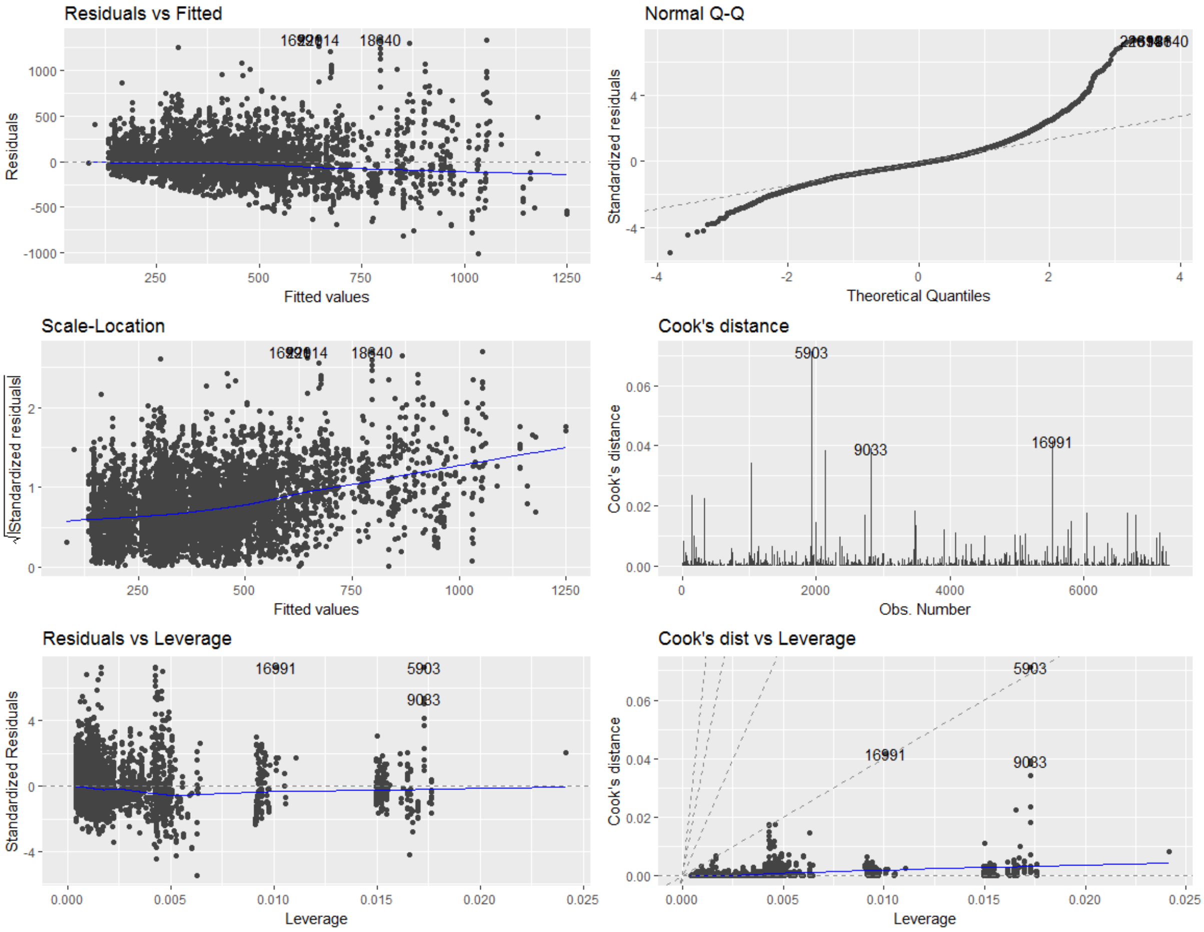
This model shows a residual versus fitted plot which has points scattered around the blue line, indicating a linear fit in the data. There are no clear curves or patterns present, supporting the hypothesis that the dependent variable has a linear relationship with its explanatory variables.

In addition, the scale-location plot has residuals spread equally along the range of predictors, which suggests homoscedasticity. Next, we see a residual QQ plot which helps to determine if the residuals are normally distributed. The plot shows residuals which have heavy tails falling outside of the expected line of a normal distribution. Due to the nature of the data being predicted (consumption), the data naturally has heavy tails. There were flags included by Thames Water to indicate the presence of leakage, which traditionally causes large right hand tails in the data, however it is not possible to exclude or distinguish between losses and leakage in all observations. Therefore, although this assumption is not perfectly satisfied, the majority of the data points fall on the normal line, and would produce sufficient results for the task in question.

The detection of correlation between the explanatory variables was handled separately in the variable selection stage, when looking for parameters to include using the manual work optimisation basket. This ensured that no two variables included within the model were correlated, so this requirement has been satisfied.

In the same way, diagnostic plots were produced for the London model, shown in Figure 15.

Figure 15 Model diagnostics for London models



The same problem of heavy tails in the residual data is present in the London data, but the same justification applies. Equally, the test of linearity is satisfied in the residuals versus fitted plots, with only three points potentially identified as unusual based on the Cook's distance metric.

12 Model testing and validation

The final household model is comprised of coefficients derived using a multiple linear regression procedure, using all data from the London and Thames Valley data sets, respectively.

Testing the robustness of the models is important and a vital step in regression modelling. There are different types of model validation which has been applied to the models in question. Largely, these are:

- Coefficient resampling, which aims to rebuild the linear models a large number of times, using a small subset of the total data available. This allows us to determine if the model coefficients are highly dictated by the sample of points used to build it by analysing the coefficient distributions.

- K-fold cross validation – This process splits the original sample size into k subsamples (here, this has been done by geographical area) and uses k-1 samples as the training data for the model build. The remaining kth sample is used as the subsequent test set in which predictions of consumption are made.

Coefficient resampling

The models in question were tested by building the model 1,000 times using 10% and 50% of the total sample size.

The stability of the R constructed model coefficients is shown in Table 29 and Table 30, which present the scale of the variance on coefficients with random samples.

It is reassuring to see, that even with only 10% of the total sample being used to build the model, that the mean and median coefficients for each parameter are very close. Further to this, the values of the coefficients are very similar to the values selected in the final model, which are presented in Table 28.

This suggests that the model is robust to the distribution of data being used to generate it, and is not being dominated by large values. This result is further corroborated with the results in Table 28.

Table 29 Variance of coefficients from 10% sample size

	(Intercept)	NUM_ADULTS	NUM_CHILDREN	EthnicHouseCat0 2	EthnicHouseCat0 3	EthnicHouseCat0 4	EthnicHouseCat0 5	EthnicHouseCat1 1	EthnicHouseCat1 2	EthnicHouseCat1 3	EthnicHouseCat1 4	EthnicHouseCat1 5	RATEABLE_VALUE
proportionate SD	0.590	0.089	0.117	0.890	1.660	0.448	1.303	0.442	0.738	0.244	1.018	31.041	0.260
median	51.47	102.82	74.35	-20.60	-10.74	-47.02	-18.00	256.38	84.42	320.95	-60.16	-2.51	0.28
mean	51.04	102.56	73.85	-20.92	-11.25	-47.28	-17.54	264.67	85.85	322.51	-59.77	2.45	0.28
Sdev	30.12	9.08	8.65	18.62	18.68	21.19	22.86	116.92	63.39	78.65	60.84	75.97	0.07

Table 30 Variance of coefficients from 50% sample size

	(Intercept)	NUM_ADULTS	NUM_CHILDREN	EthnicHouseCat0 2	EthnicHouseCat0 3	EthnicHouseCat0 4	EthnicHouseCat0 5	EthnicHouseCat1 1	EthnicHouseCat1 2	EthnicHouseCat1 3	EthnicHouseCat1 4	EthnicHouseCat1 5	RATEABLE_VALUE
proportior	0.195	0.030	0.039	0.296	0.574	0.152	0.414	0.137	0.233	0.076	0.327	52.430	0.091
median	52.14	102.72	73.57	-21.16	-10.75	-47.80	-17.82	263.00	88.62	315.07	-59.55	0.20	0.27
mean	51.96	102.74	73.64	-21.11	-10.96	-47.62	-17.86	261.64	88.77	316.57	-59.55	-0.48	0.27
Sdev	10.14	3.07	2.90	6.25	6.29	7.25	7.40	35.76	20.71	24.19	19.49	24.98	0.02

12.1 Testing seasonal effects

In addition to the resampling and cross validation described previously, tests were conducted to determine if a better model would result from combining the results from models built using summer and winter data separately. This is a somewhat unconventional test, but alleviates questions which could arrive surrounding the models ability to deal with seasonal changes.

Full details of this analysis can be found in the Appendix section 22.1

12.2 Internal validation of the models

12.3 Resampling

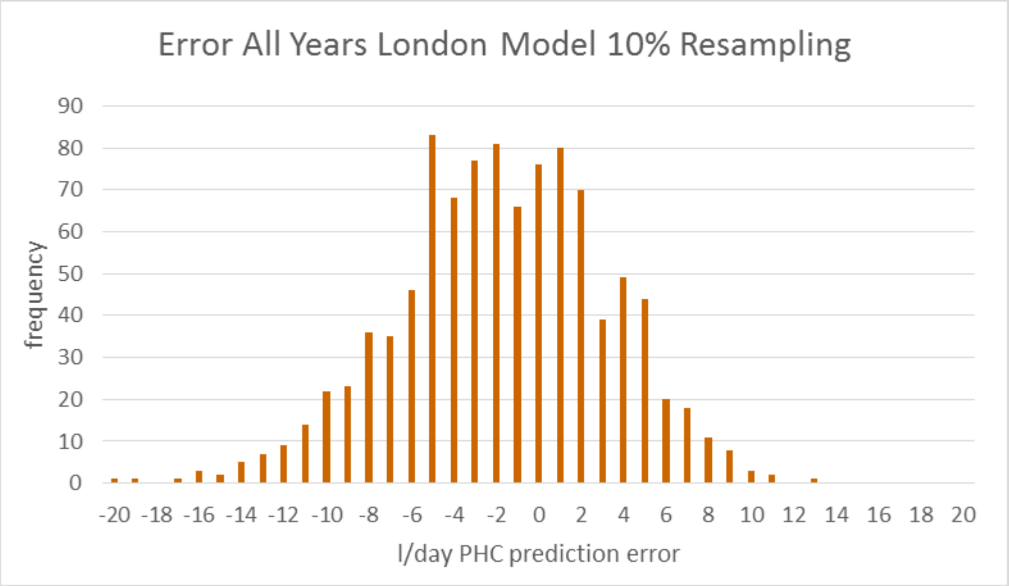
Having established the robustness of the base model, the sample variance for London and Thames Valley models were subjected to testing as follows:

Two models were derived; one using all data, the other using 2007 data alone. It was decided to use all years' data to mitigate annual variance due to climatic factors. Both models were used to predict one thousand separate 10% samples of model data.

The individual predicted household consumptions were summed to represent a zonal consumption estimate, and the error calculated.

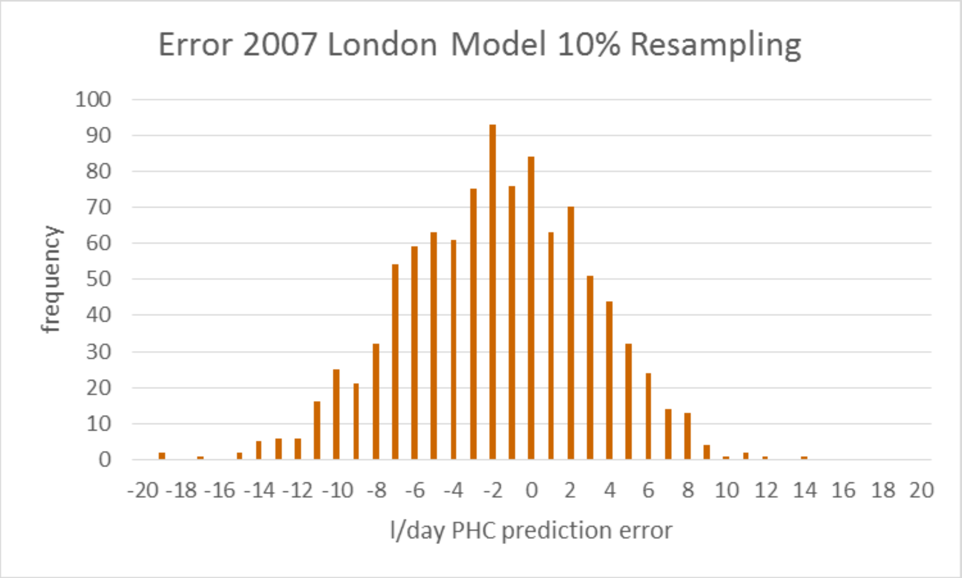
The similarity of the results shown in Figure 16 and Figure 17 show that the performance of the model is similar whether the prediction sample comprises the model trainer set or not, and that a model built from 2007 data is equally effective in predicting subsequent years consumption. Error is due to variance of the 10% population sample.

Figure 16 Prediction error on 10% of London data set using all data model



A model built from 2007 data slightly outperforms a model built from all years, although errors are very small in either case with 4l/day roughly equating 1% error.

Figure 17 Prediction error on 10% of London data set using 2007 data model



12.4 Independent trainer set testing

With time series data of variable quality in the data set, an opportunity to create a model prediction test with parallels to forecasting presented itself.

In this independent test, the base year (2007) was used to predict subsequent year’s consumptions in aggregate (2008 onward).

Figure 18 Independent London consumption forecast prediction

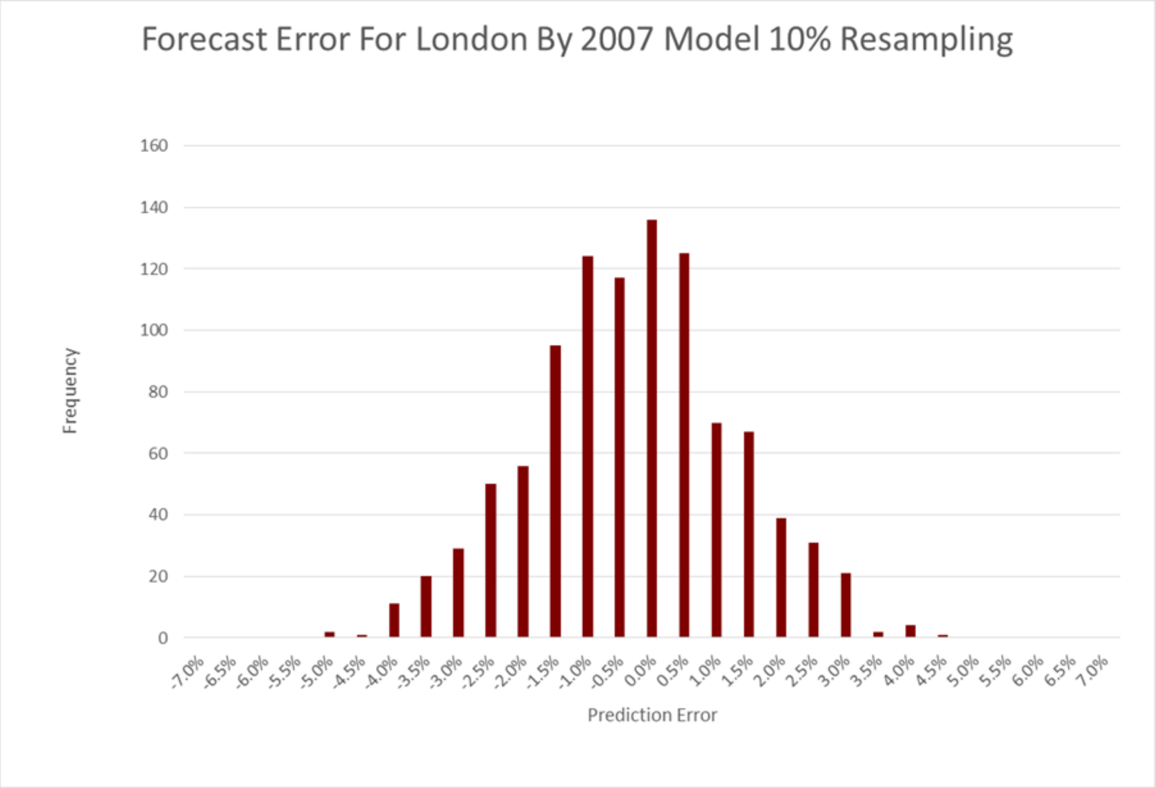
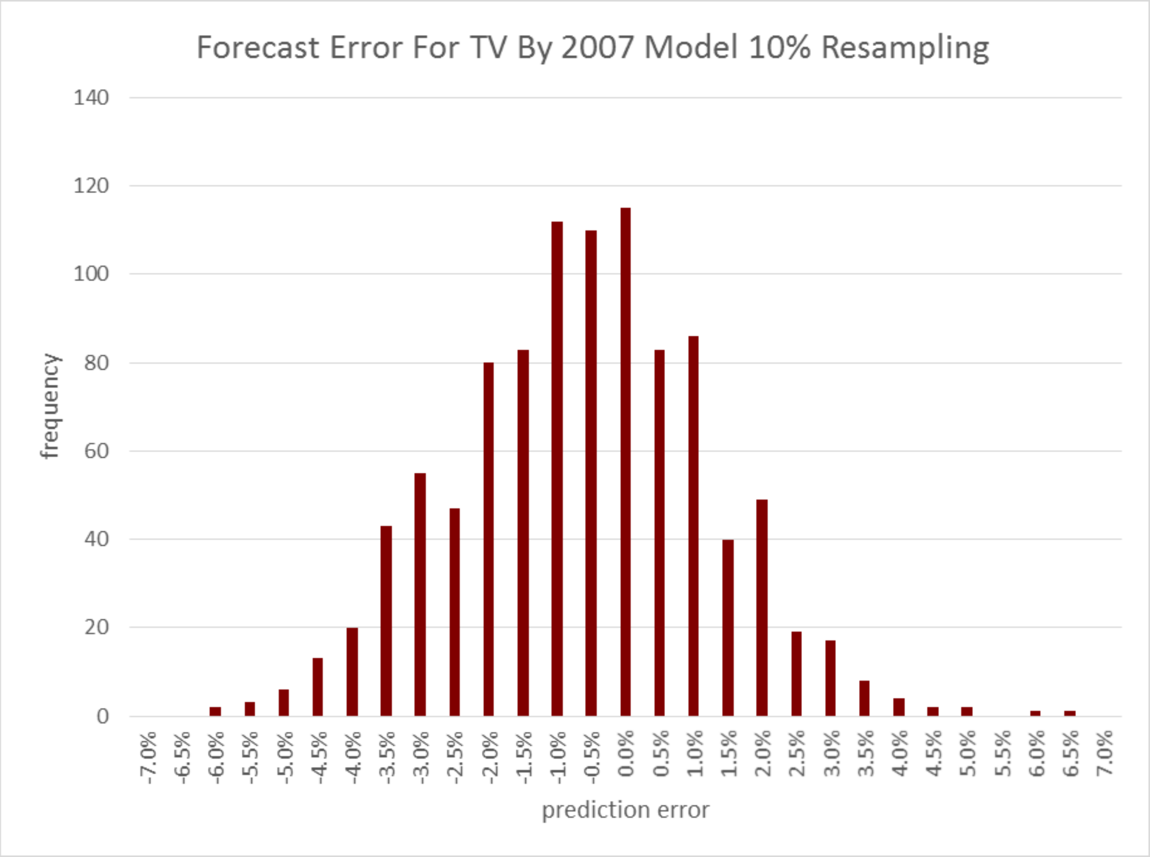


Figure 19 Independent Thames Valley consumption forecast prediction



The standard deviation of the independent prediction test error is 1.57% for London and 1.86% for Thames Valley (Table 31).

The mean error has components of random error, base year (2007) bias, as well as consumption trend.

Table 31 Independent forecast test statistics

Forecast Error Stats 10% Sample		
	London	Thames Valley
Mean	-0.55%	-0.87%
Standard Error	0.05%	0.06%
Median	-0.52%	-0.81%
Mode	0.00%	0.00%
Standard Deviation	1.57%	1.86%

Forecast Error Stats 10% Sample		
	London	Thames Valley
Sample Variance	0.02%	0.03%
Kurtosis	-0.074	0.095
Skewness	-0.003	0.035
Range	9.87%	12.45%
Minimum	-5.47%	-6.35%
Maximum	4.40%	6.09%
Count	1000	1000
Confidence Level (95 %)	0.10%	0.12%

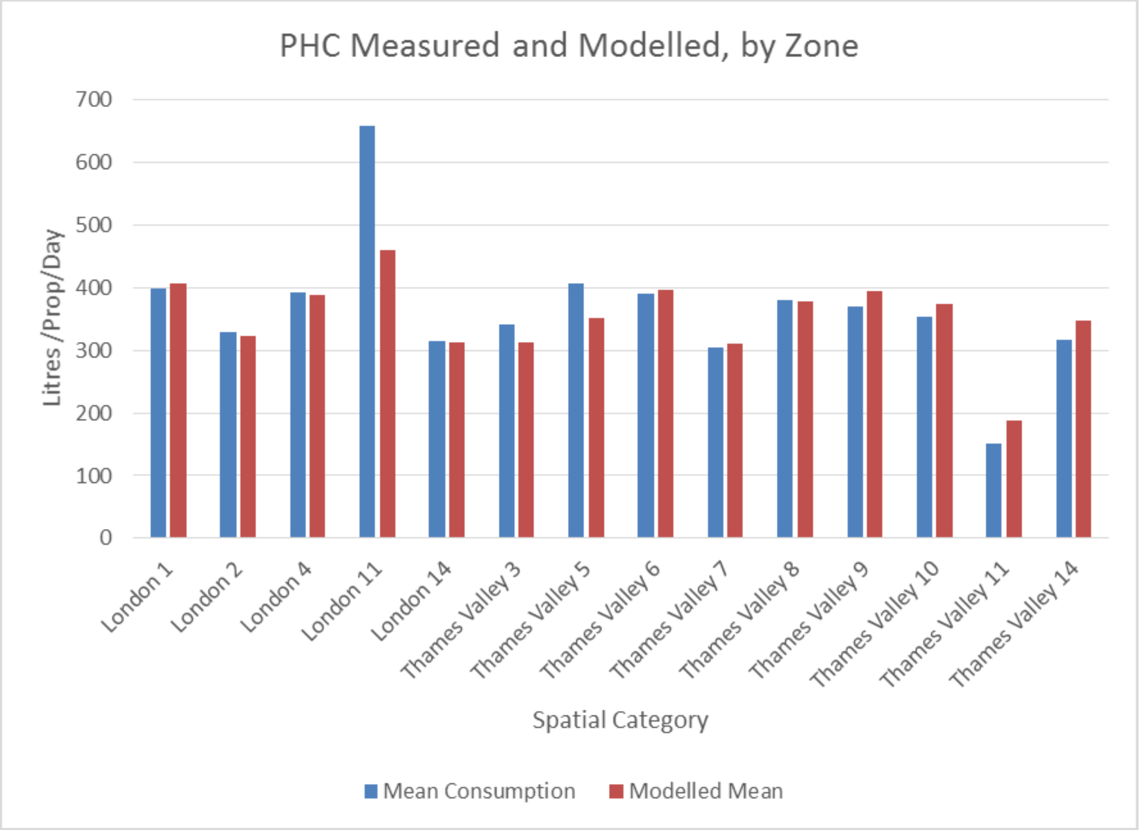
These tests demonstrate that the models used are adequate for predicting 10% of the DWUS data set consumptions to better than 5% at 95% confidence. The independent test shows that over 99% of results are within the 5% error bound. The next set of tests looks at the spatial application of the model within the DWUS data set.

12.5 Cross validation by area

One of the key performance indicators of a consumption model is its ability to predict consumption using areas which have been removed from the trainer set. This is called cross- validation, and is usually performed by removing ‘k’ groups of equal sizes from the trainer set in sequence, and using these omitted groups to predict consumption.

As a slight variation on this method, and to test the model spatially, the k groups were defined by the Thames resource zones, as well as London and Thames Valley area giving 14 classified areas. Each model was then used to predict the consumption in the excluded area. Figure 20 shows the difference between the modelled and the actual mean consumption values for all zones using this technique.

Figure 20 Internal spatial testing



The results show that the modelled mean consumption value is very close with the actual consumption, meaning that the models are able to predict new data very well in all zones. The only zone which has a mean consumption much higher than the modelled is 'London 11'.

Using, Table 32, which shows both the modelled and mean consumption values for all zones, as well as the sample size of the area in which predictions were made, it can be seen that the sample size for four of the 14 zones are too small to allow valid results to be obtained. Therefore, 'London 11', 'London 14', 'Thames Valley 9' and 'Thames Valley 11' can be omitted from the analysis.

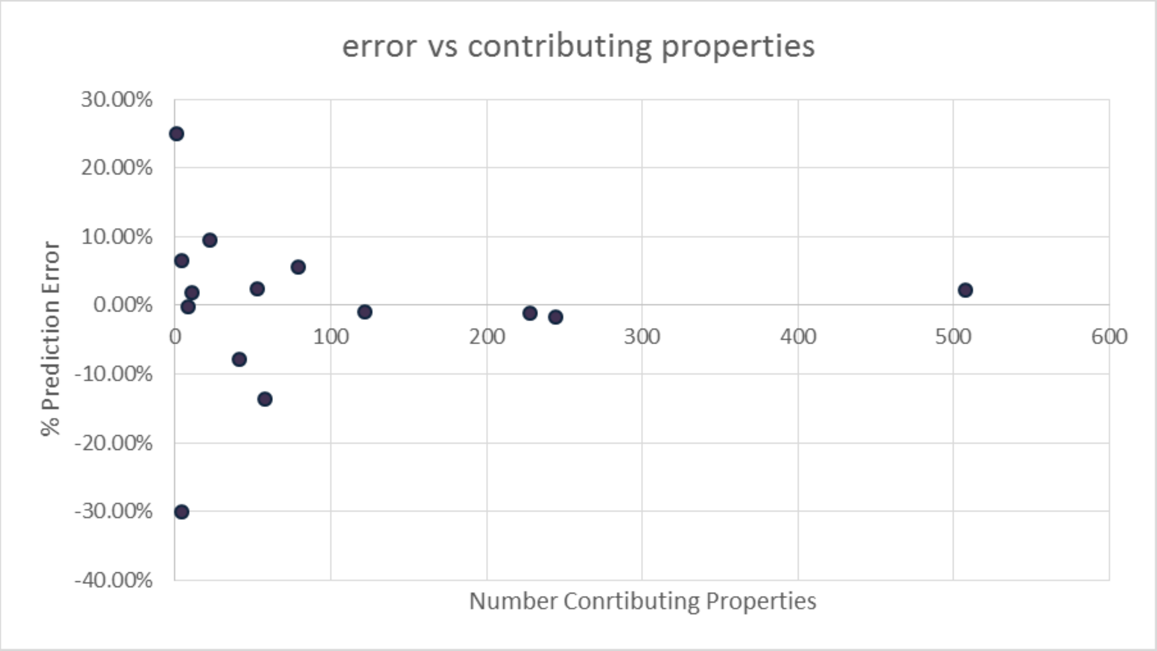
Table 32 Zonal property sample size for spatial testing

Spatial Category	Mean Consumption	Modelled Mean	Sample size for prediction
London 1	397	406.6	508
London 2	329	323.1	244
London 4	393	388	228
London 11	657	459.9	-
London 14	314	313.2	-
Thames Valley 3	340	313	41
Thames Valley 5	406	351	58

Spatial Category	Mean Consumption	Modelled Mean	Sample size for prediction
Thames Valley 6	389	396	11
Thames Valley 7	304	311	53
Thames Valley 8	381	377	122
Thames Valley 9	370	394	-
Thames Valley 10	353	373	79
Thames Valley 11	151	188	-
Thames Valley 14	317	347	22

Using this table, we can plot the sample size against the prediction error for the 10 valid zones.

Figure 21 Internal model error vs sample size



These results indicate that the error reduces as the number of samples increases, however all zones have an error of less than 10%, which is a reasonable figure. This shows that the model is able to predict previously unseen zones well. These errors are before trend and climatic adjustments have been added.

13 Aggregate to zonal level

The model which has been built thus far has used household data to derive an equation for predicting household level consumption. However, the model will be used zonally and so the interpretation of the household coefficients is slightly different.

Due to data constraints and time limitations, it is not possible to derive a household forecast for each property for every zone to determine zonal consumption. Therefore, to use the model at this higher level, aggregate numbers are used where previously household data would be placed. For example, instead of household adult and children numbers, the total number of adults and children per zone are multiplied by their respective coefficients.

In a similar way, the IBP flags will no longer be binary, but will include the number of properties in each category to be applied to each coefficient.

This is a very subtle detail which explains the jump from the household to the zonal predictions. However, it may now be clearer as to the importance that the parameter selection accounted for the possibility that the selected parameters were forecastable at a zonal level.

By aggregating the model in this way, we are able to consider other parameters which weren't possible to consider at household level. For example, weather effects are not easily distinguishable at such a fine resolution, but can be considered zonally.

13.1 Developing a metered property coefficient

Before searching for other variables to help explain zonal variation, and since the original model was built using unmeasured households, investigations are conducted to determine the suitability of the model when predicting *measured* consumption.

The property segmentation required in the forecast is between measured and unmeasured properties. The majority of the DWUS dataset is unmeasured (previously shown in Table 14), but the dataset also allows the examination of the difference between measured and unmeasured consumption and the transition between the two. Greater than usual insight is afforded because the difference is contextualised by the parameters in the model. The use of these properties to derive a metered coefficient must be done carefully, since the sample may be considered biased due to the nature of the study. However, these households represented the few measured households available, in which the transition from unmeasured to measured status has been tracked.

A measured coefficient can be derived directly by summarising the data to quantify differences using different stratifications, and measured status can also be used as a binary variable within the linear model to compensate for different population trends. A more sophisticated modelling approach is to quantify the difference between predicted and observed consumptions for the measured population, based on the unmeasured model.

The filtered data set counts just legitimate consumption, but consumption including leakage and wastage can also be examined by partitioning the data set with the relevant error flag. Data with leakage must also be scrutinised to evaluate whether removal of this data skews the modelling of legitimate consumption. In other words, it might be expected that properties with high wastage and leakage are also more profligate with legitimate consumption as defined by the Thames auto validation process.

The meter coefficient was separately derived for London and Thames Valley by predicting measured consumption for both summer and winter, using summer and winter derived models and defining the difference for both summer and winter as the effects of measured billing. It has been shown (section 22.1) that averaging coefficients for summer and winter produce a coefficient very close to an annually derived one.

Table 33 presents how the base unmeasured model was used to predict measured consumption. The difference between the estimated consumption and the observed consumption is the difference that we attribute to measured billing.

The reason for splitting the derivation seasonally was to address the possibility that measured and unmeasured difference has a large seasonal component. This is shown Table 33 with the summer adjustment being slightly larger for Thames Valley and London.

The derived value was sense checked by running a version of the model with all properties and a binary dummy variable for measured status, as well as conducting a direct comparison of measured and unmeasured consumptions baselined by property type and occupancy.

The effectiveness of the meter coefficient was tested against zonal reported consumptions using both the original additive coefficient and a multiplicative version. The additive coefficient produced a better R-squared at prediction, (which was established at the zonal calibration level) and was therefore selected.

The first iteration linear adjustment evolved into the dynamic version once the strong relationship between the correction required to the meter coefficient and zonal meter penetration was observed. The unmeasured model is unaffected.

Table 33 Measured consumption prediction from unmeasured model (l/prop/day)

Measured Prediction Thames Valley Winter	Median	Mean
Unmeasured values	333	362
Measured values	227	255
predicted measured values	346	351
Metered Supplement	-119	-96
Measured Prediction Thames Valley Summer	Median	Mean
Unmeasured values	333	362
Measured values	233	260
predicted measured values	347	366
Metered Supplement	-114	-106
Mean Thames Valley metered supplement		**-101 l/p/day**
Mean London metered supplement		**-61 l/p/day**

Measured Prediction London Winter	Median	Mean
Unmeasured values	329	376
Measured values	186	239
predicted measured values	309	294
Metered Supplement	-123	-55
Measured Prediction London Summer	Median	Mean
Unmeasured values	342	395
Measured values	194	238
predicted measured values	316	305
Metered Supplement	-122	-67

13.2 Meter optants and smart metering effects

As well as accounting for measured properties within the model, the effect of optants and smart metered households are to be included.

A number of optants are added each year in the Thames Water forecast. Data analysis has shown that newly optant properties have a lower occupancy, and hence consumption, than an average measured property. Because of the churn rate in properties, we can deduce that the optant status does not last indefinitely, and the household will move into existing measured status as new occupants who have not opted for a meter on that property replace the optants.

Many measured properties were once optant properties, and introducing a coefficient for lower optant consumption is difficult because the average of measured properties contains optant properties also, and with no clear cut-off point the distinction cannot be made.

A modelling assumption is that optant type customers are present in the existent measured population in a representative proportion. In other words, the optant category represents purely optant, whereas the measured category is a mixed category including optants, ex-optants, would be optants, and “noptants” (never optants). Any of these categories may move into a pre-metered house and not have the chance to opt.

We therefore make the assumption that optants will be added to the measured population at annual average measured occupancy which approaches mean as meter penetration approaches 100%. Although this may not be true of optants as they enter measured category, it preserves a relative dynamic equilibrium in measured occupancy, (to the exclusion of changes due to meter penetration and overall occupancy changes).

Progressive metering is a new category where smart meters are fitted without optancy. Artesia have been advised by Thames Water that smart metering will prompt a 12% saving over standard metering. In other respects, progressive metering has been treated the same as measured, except the occupancy is assumed to be the mean of unmeasured. There is facility in the model dashboard to change the 12% saving to whatever is required.

14 Zonal adjustments

Due to the dates in which the data was collected, the demographic data included within the model is weighted to FY 2007 and can be viewed as the model base year. The further we move from the base year, the more the variables may depart from their original values, and the model may have reduced predictive power.

Temporal validation is further complicated by trend in un-modelled elements (such as the introduction of low volume cisterns), and variations in demand due to weather variations. This creates a time series trend in the residuals overlain with an apparently random variance which we attempt to explain with weather parameters. This process is termed residual analysis.

14.1 Weather and trend modelling

Internal residual analysis was carried out on the base model. This compares predicted and measured consumptions using historic data. Because year 2007 is the effective base year, the trend was calculated from this starting point.

The proportionate difference in modelled and measured consumptions for the DWUS data in time series was defined as the observed trend for London and Thames Valley. This was calculated both for measured and unmeasured populations.

The observed trend was strong in Thames Valley, and weaker in London. Once this was adjusted for the observed trend, residuals were correlated against annual and seasonal weather data from Heathrow weather station.

Figure 22 shows a strong relationship between summer residuals and rainfall; Figure 23 shows a somewhat weaker relationship on an annual basis.

This disparity is to be expected as exterior water use is the main reason for these adjustments to consumption, and is stronger in the summer.

The scale of the calculated coefficients for the summer and annual analysis is consistent with this division, as inclusion of the whole year would both damp response to weather variables and introduce more uncertainty, and this we see. It is a strong validation of the model that nearly half the residual error is attributable to rainfall effects.

Figure 22 Summer residuals vs rainfall plot

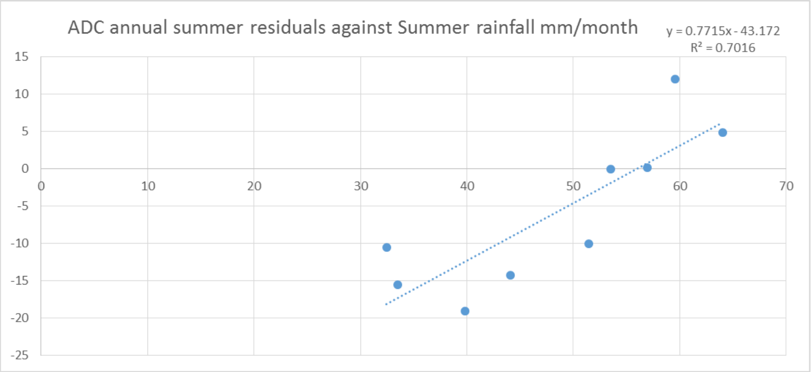
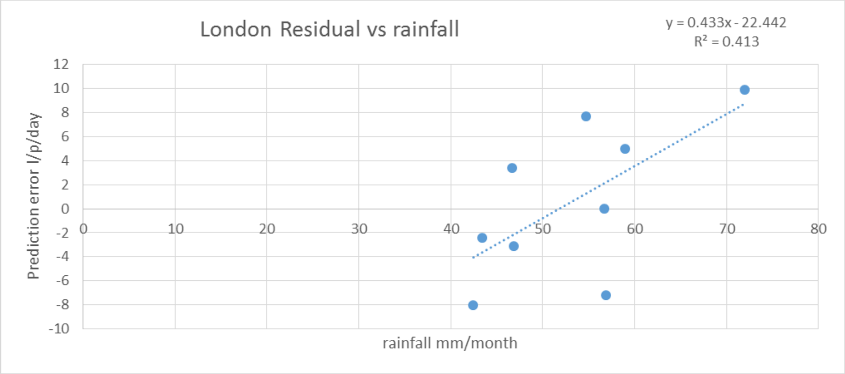


Figure 23 Annual residuals vs rainfall plot



The residuals were then adjusted for observed trend and weather influences to obtain a residual error unaccounted for by any stage of the modelling process (Figure 24, Figure 25). The adjustments place an error envelope of around 12 l/p/day around the forecast from base year 2007. This includes parameter drift, and equates to around 3% of PHC.

Figure 24 London model residuals

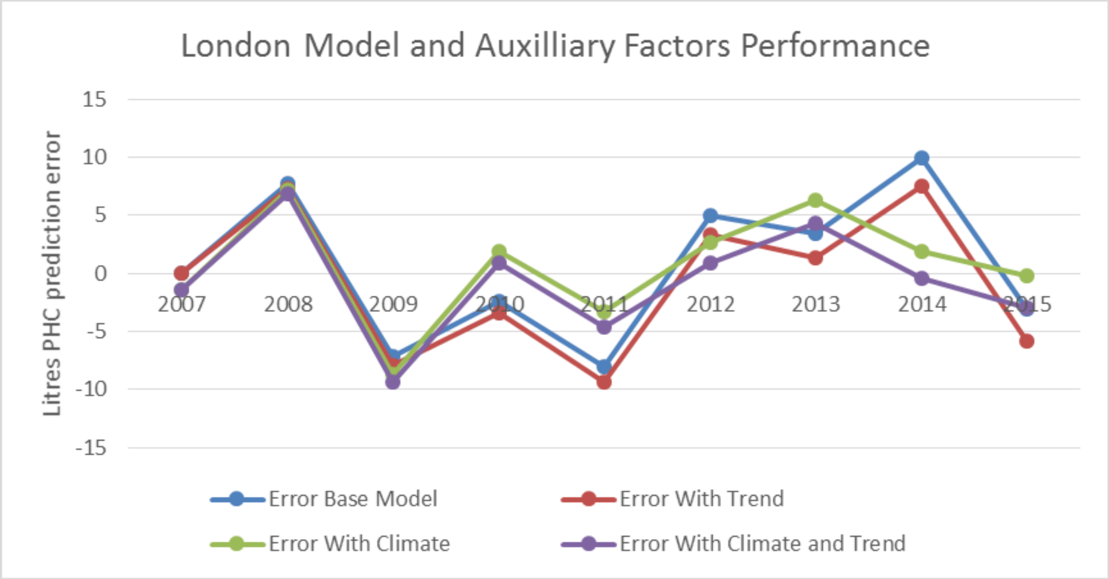
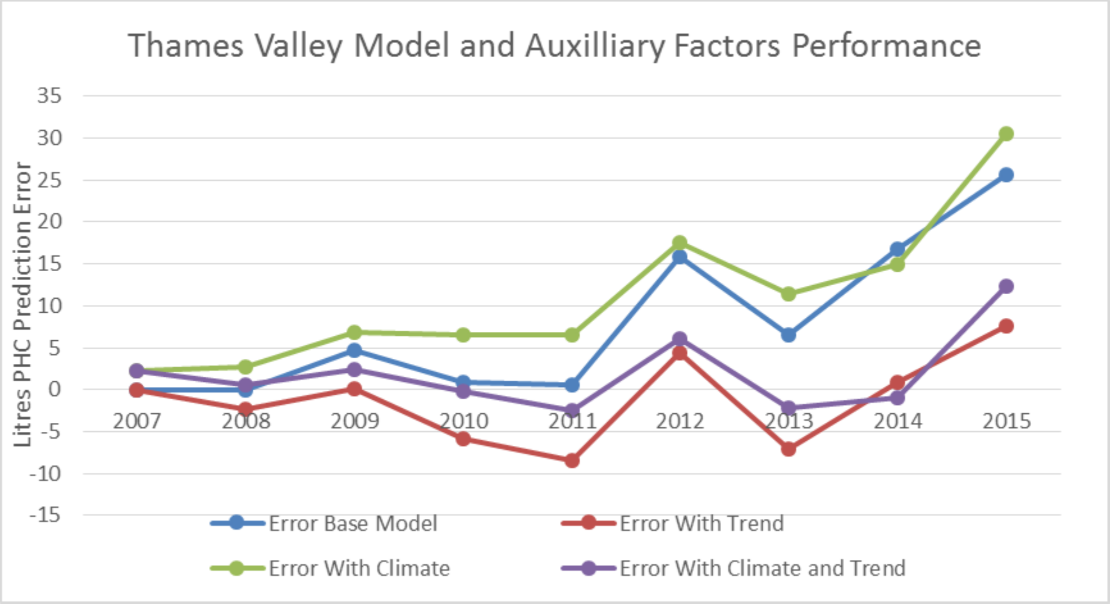


Figure 25 Thames Valley model residuals



14.2 Base year normalisation

The base year for calibration is financial year (FY) 2015/16; selected as the most recent year for which a full and recent data set was available in terms of both reported consumptions and associated demographics. There is no recent database specifically quantifying micro-component consumptions for the Thames Water area.

If the model is to correctly forecast a 'normal' year, then the base year consumption must also be normal. The approach taken to achieve this is bootstrapping, and relies on a second level of linear modelling.

Firstly, the uncalibrated temporal residuals are modelled against weather parameters using a secondary linear model. The model is built by first selecting the most appropriate variable from a selection including temperature, rainfall and sunshine hours. Once the variables have been selected and the model built, the resultant weather-consumption model is applied to the base year weather parameters to produce a base year consumption correction for a normal year, which is fed back into the model for base year calibration on future projections.

In the case of the London model the weather normalisation is 2.5 l/p/day, which approximates to 0.5%. This is effectively a normal year as 0.5% is within modelling error we'd expect.

15 Model calibration

So far, the model has been validated at DWUS household level individually or in aggregate spatially or temporally by way of resampling and cross-validation. However, since it will be used at zonal level, and amendments have been made for measured households, the results will need to be calibrated against reported consumptions in six WRZs, the largest being London with 83% of properties and 84% of consumption.

There are a number of reasons why the consumption model is expected to under-predict reported consumption.

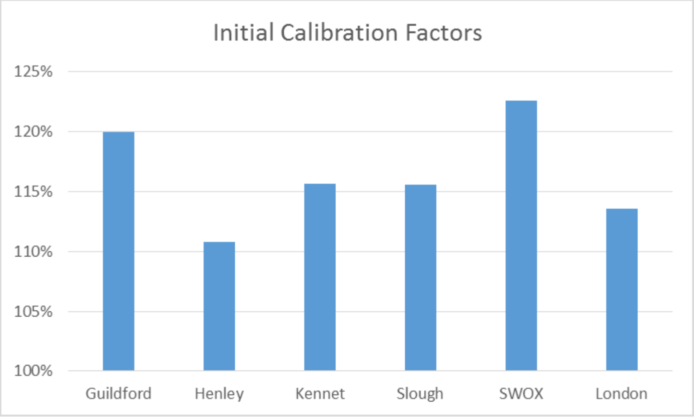
- The trainer data was principally unmeasured but metered. Metering raises awareness of losses and water use behaviour and may cause the model to under-predict the zonal consumption.
- The sample was self-selecting of a demographic willing to subject themselves to scrutiny. This would exclude for instance properties that house illegal tenants which typically have high occupancy. Properties with high losses are also likely to be circumspect about such scrutiny.
- The trainer set excluded properties with high losses.
- The transient population are likely to be underrepresented in a long-term study. It is not known how this would impact on mean consumption.
- Measured consumptions are not corrected for MUR

15.1 Measured and unmeasured households

Firstly, unmeasured calibration was carried out to determine a zonal calibration factor. Figure 26 shows the magnitude of these factors. The difference between the measured and unmeasured consumption is given by the additive factor in section 13.1.

Correction factors of between 10 and 23% need be applied to the raw unmeasured outputs. A large proportion of this is likely due to the exclusion of properties showing leakage, as per the bullet points above.

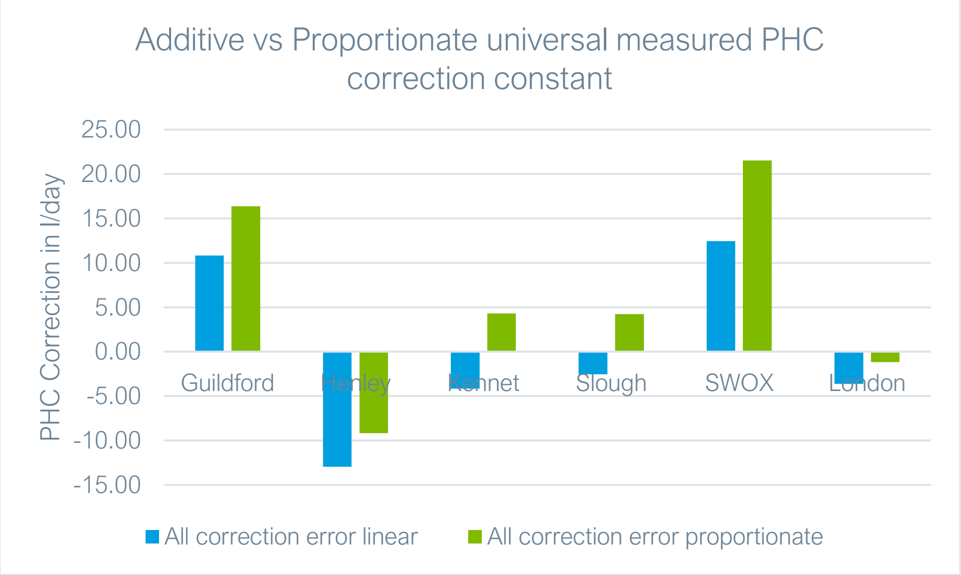
Figure 26 Zonal unmeasured HH calibration factors



The meter coefficient is supposed to express the difference between measured and unmeasured properties, and the correction factor required to the meter coefficient to bring the measured modelled proportionately in line with the unmeasured (i.e. requiring the same zonal correction factor for both measured and unmeasured shown in Figure 26) was determined.

The mean zonal meter coefficient correction factor produced a smaller range of error as an additive factor than a proportionate factor (Figure 27) if a single universal factor was applied to all zones. This lends weight to the use of an additive rather than proportionate meter coefficient.

Figure 27 Additive vs Proportionate meter coefficient correction factor performance



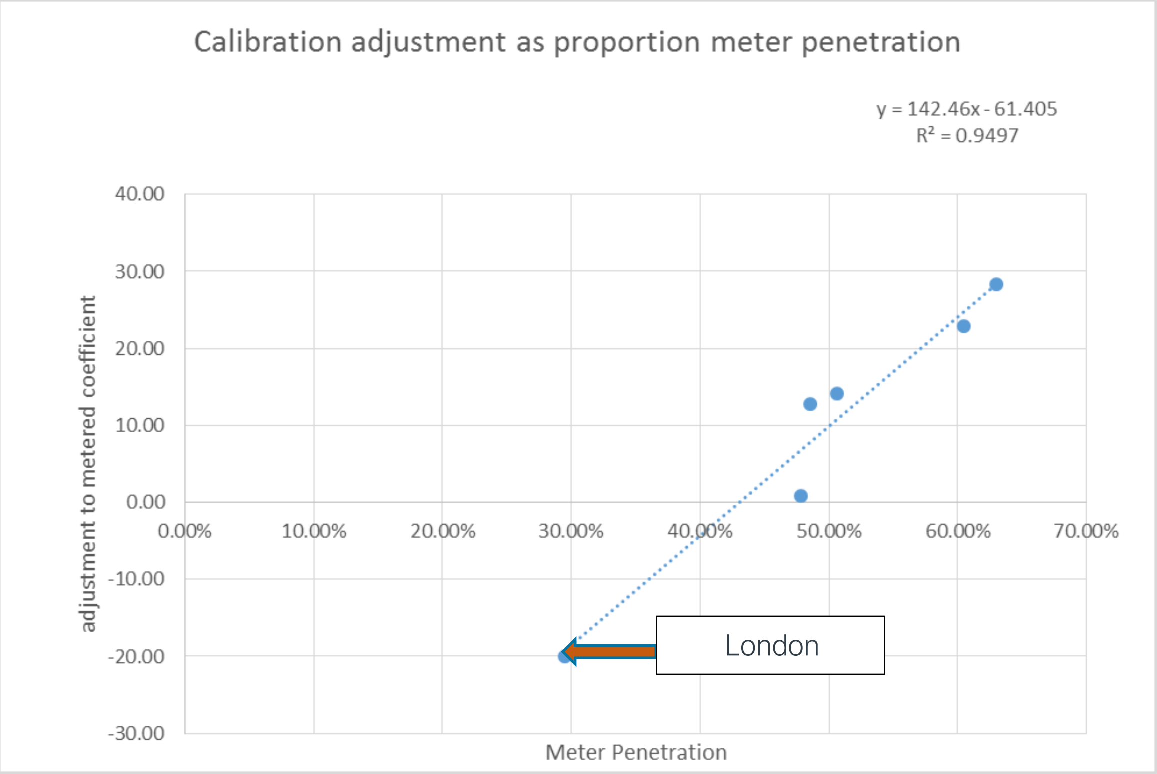
This residual analysis left the question as to why the zonal level calibration had a much greater variance than would have been expected from the internal model spatial validation (Section 12.5) considering the much larger sample prediction size.

There is greater correction required of the measured model, and that being the case it was logical that the residuals may be explained by a zonal characteristic related to metering. The obvious candidate being meter penetration.

The adjustment required to the zonal meter coefficient was plotted against zonal meter penetration. The objective was to establish a relationship between the meter coefficients required to bring the measured model in line with the unmeasured model by zone.

The unmeasured model also shows a variance in accuracy, but this is being regarded as a zonal issue and we are not attempting at this stage to close this gap with the meter coefficient. Indeed, the unmeasured zonal calibration factor shows no significant relationship to meter penetration with an R-squared value of 3%.

Figure 28 Determination of metered coefficient correction by residual analysis, for TV WRZs.



The correction to metered coefficient based on meter penetration Figure 28 applied to the Thames valley coefficient gives a dynamic meter coefficient (DMC) with intercepts at 0% and 100% penetration of 162l/day and 19.54l/day.

In order to support this finding, the small sample of DWUS properties that transitioned from unmeasured to measured was studied in the context of the rest of the DWUS population.

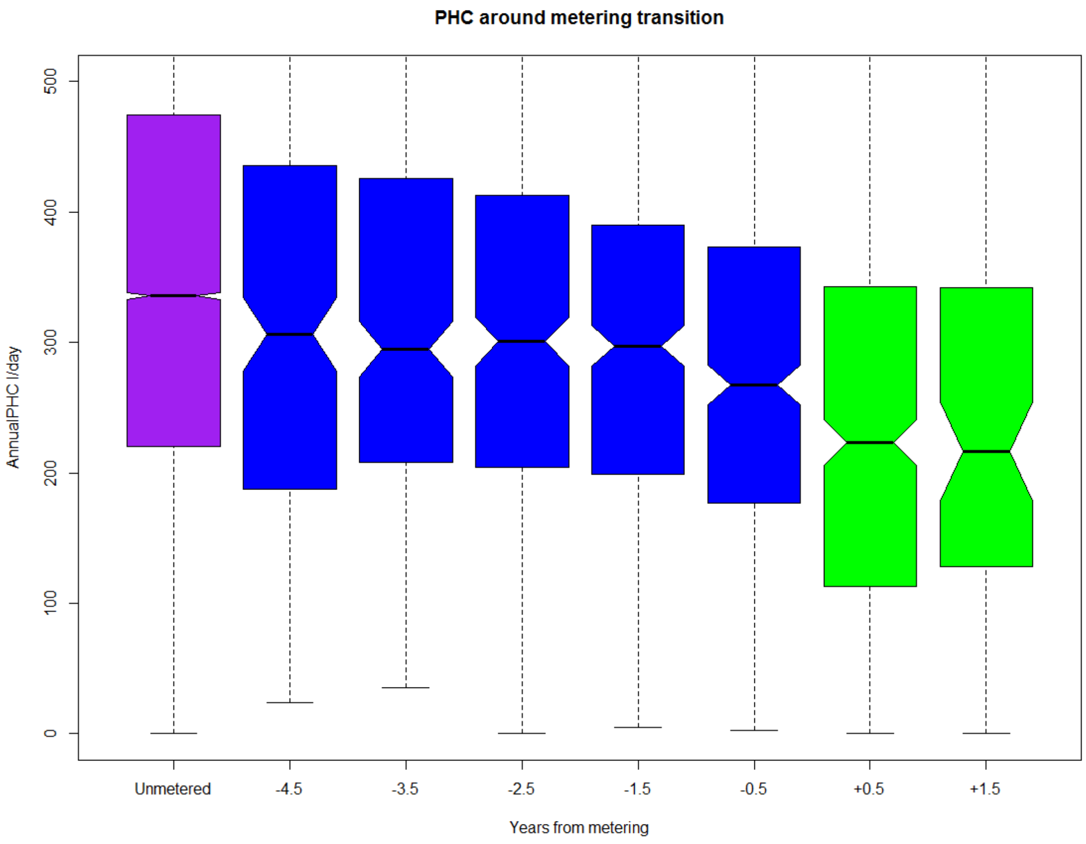
15.2 Calibration by transition analysis

The observed shift in the metering coefficient with meter penetration is the manifestation of a predicted theoretical outcome concerning populations.

It is assumed that customers who believe themselves to have low consumption are more likely to opt than the average consumer.

Figure 29 shows median and distribution of PHC in the optant DWUS sample for the unmetered population (purple), optants as they approach the metering point (blue) and post metering population (green). The notches represent standard error around the median.

Figure 29 Box plot PHC of unmetered and transiting to metered households



PHC starts to decline significantly in the year prior to metering, but the pre optant population are already significantly lower consumers than the unmeasured population. In this case, the difference in PHC across the optancy process is in the region of 70 l/prop/day in the median PHC at optancy minus 1.5 years through optancy plus 1.5 years.

The difference between unmeasured PHC and post optancy PHC is near 100 l/prop/day PHC. These figures represent a sense check of the meter coefficients calculated in section 13.1 at the existing Thames area meter penetration of circa 40%, and also demonstrates that pre optant properties are closer to unmeasured than post optant.

In terms of strategy for implementing water efficiency savings, the implication is that reduction in consumption due to measured billing is more significant than behavioural/ demographic/ideological differences between pre optants and unmeasured populations.

15.3 Applying the dynamic meter coefficient

The difference between measured and unmeasured household consumption for identical model demographics is large at current meter penetrations due to preferential meter uptake by low consumption customers. The mean zonal difference is 145 l/property/day at base year, although a proportion of that is accounted for by model inputs such as occupancy and property type.

As meter penetration approaches 100% the consumption characteristics of the measured population approach the population average, and the measured/unmeasured difference approaches a value which equates to the average water saving efforts by measured householders. The relationship between measured and unmeasured consumption will be altered dependent on metering motive. Where metering is not optant (i.e. new builds, progressive), we might expect unmeasured consumption to remain relatively constant because there will not be the preferential removal of low consumption properties. Where optancy drives metering we expect average PHC of the unmeasured population to rise.

In an effort to find a reasonable intercept for a dynamic meter coefficient (DMC), we assume the figure we require for DMC at 100% metering represents the underlying water saving due to measured billing as distinct from the demographic difference between measured and unmeasured populations. It is thought from the more normal distribution of measured consumptions that this saving is chiefly due to reduced plumbing losses in measured households, as it has been demonstrated that losses are very skewed and well represented by a lognormal distribution (Figure 31, Figure 3).

The saving has been estimated by Artesia to be in the range of 20 to 40 l/prop/day based on transitioning properties adjusted for occupancy changes. This figure is impossible to calculate from existing consumption data because the non-consumption component classified as supply pipe leakage cannot be separated from internal plumbing losses.

The intercept at 100% metering, can be estimated, but the route taken will vary depending on metering motive.

15.4 Double dynamic meter coefficient

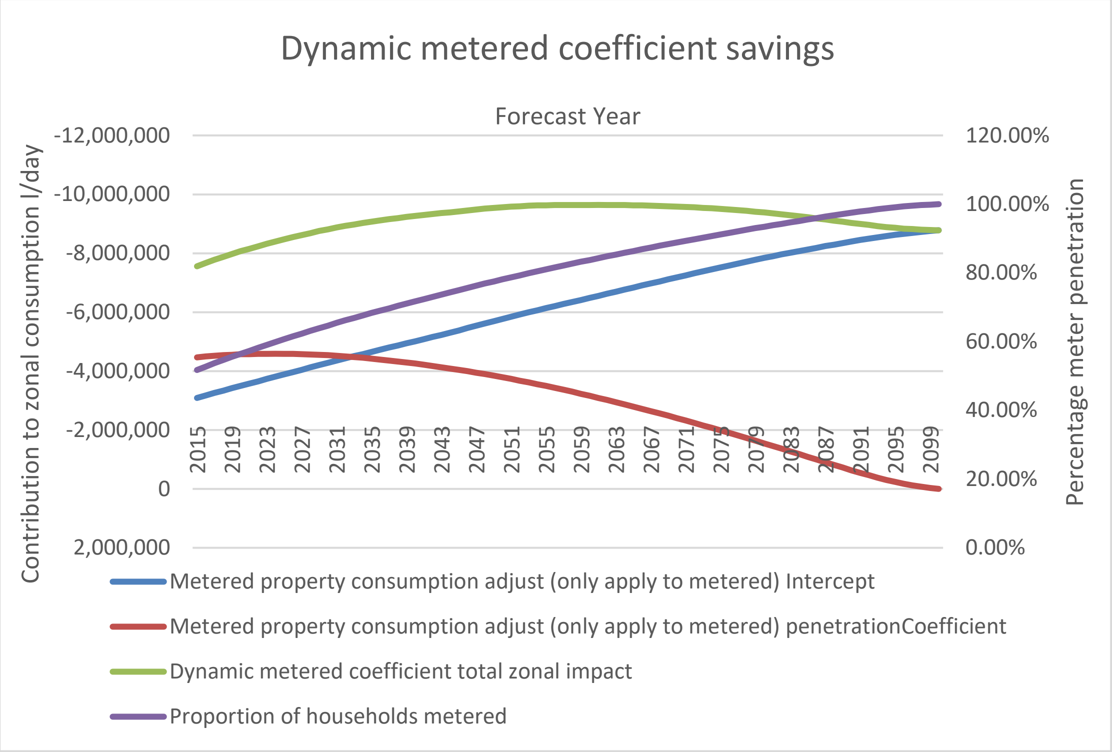
15.4.1 Model failure at high penetration

The intercept term in the DMC was justified in its original form by the fact that a purely linear relationship appeared to best fit the data. This also being based on static data (as opposed to a time-series) suggested little reason to use a more complex model.

In application, however the model runs into issues when we try to model a meter penetration upwards of (circa) 80%. As seen in the plot the consumption saving from metering begins to drop at around 82%.

From a modelling perspective, we should expect to predict a law of diminishing returns in terms of the difference between measured going forward and unmeasured at base year due to convergence of measured consumption with mean, but for the total saving in consumption to actually reduce is somewhat unrealistic. This applies to all zones at high meter penetrations.

Figure 30 Dynamic meter coefficient savings for Kennet.



In Figure 30 we see that the dynamic meter coefficient contribution to the saving diminishes quickly while the 'Intercept' contribution tracks the proportion (as of course it should). In this region the intercept term dominates the picture, and the fact that the model predicts unrealistic values at high penetration suggests that the intercept term is insufficient in this region; we therefore see justification for further refinement to the model.

15.4.2 Losses term

The reported zonal consumptions against which we calibrate model outputs include an adjustment for supply pipe leakage, a proportion of which is allowable as consumption if it is deemed to be on the customer's property. All leakage and wastage due to internal plumbing issues such as overflowing cisterns, dripping taps and customer side supply pipe leakage etc. are termed losses, and are part of billed consumption.

We now refine the model using the hypothesis that, from understanding of the transition between unmeasured and measured properties, there is significant relationship between the distribution of unmeasured consumption and high losses. Data shows that measured property consumption is distributed more normally than unmeasured properties, which tend toward a log-normal distribution. The latter is reminiscent of the distribution of household leakage, and indeed previous work has suggested this is the case (refer back to page 19, Figure 3).

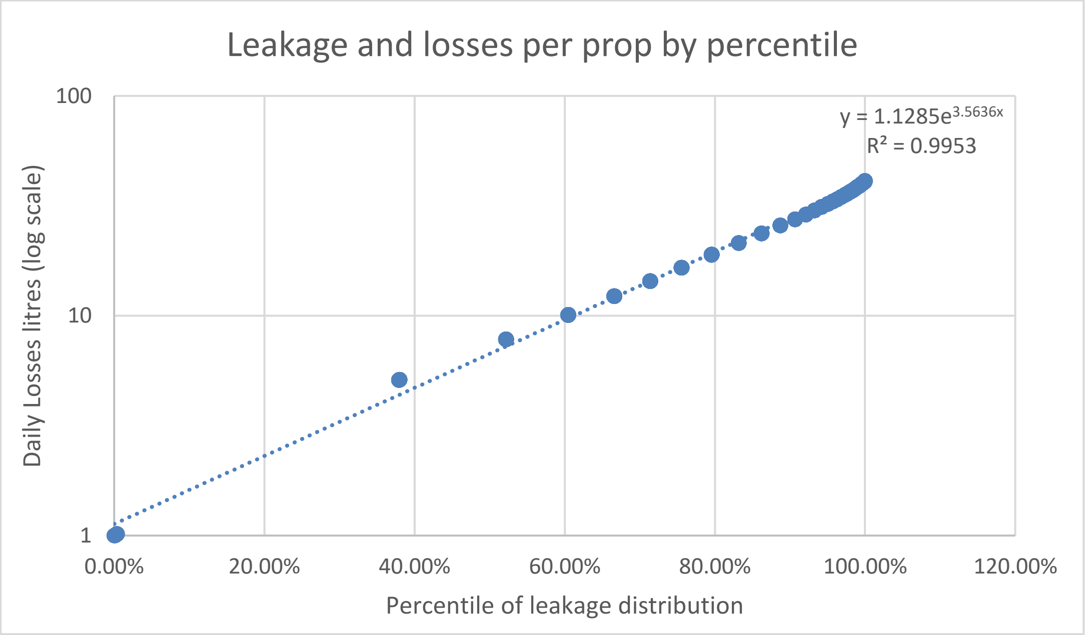
Lowest 15-minute night flows in the DWUS database were used to calculate a nightly losses rate. There was no significant all-night flow in 63.5% of the data: i.e. a minimum hour of less than 1 litre.

If we take the cumulative mean night losses against a losses percentile of the DWUS population we find a log linear relationship that takes off at 63.5% with an intercept of around 40 l/day at 100%. This is the saving that we assume we make as metering approaches 100%, if we assume that measured billing has the effect of prompting the saving of losses.

It is not credible to say that the first 63.5% of measured population have no losses to save, so the log linear relationship is rebased with the 63rd percentile moved to the zero percentile (Figure 31). This allows for some losses savings at low meter penetrations – and we have seen this in the transition from unmeasured to measured.

Rebasing the equation does raise the mean losses at a given percentile, but it must be borne in mind that low level losses may be at or below meter stall speed so these will be underestimated in general.

Figure 31 Percentile of distribution against losses level.



On this basis we refine the model on the assumption that properties are metered in a broadly but not precise increasing order of losses. This assumption allows for a mix of optancy, progressive and new property metering. The picture that DWUS data gives is that the distribution of mean losses per household by percentile can be modelled as an exponential trend to a good approximation ($R^2 = 0.9968$). From this we can introduce a term to replace the static 'savings' intercept term of the DMC which allows for the lognormal distribution of losses. Note that in Figure 31 the percentile is the percentile of properties with registered losses above 1 l/day. Registered losses below 1 l/day contribute 63.5% of data.

Due to meter under registration (MUR), losses will not register until they hit starting flow rate, often claimed to be around 20l/day. Mean supply pipe leakage is reported at between 16 and 20 l/day, so it could be argued that all billed losses can be averaged to be measured continuous flows, as mean supply pipe leakage will not register unless aggregated with a simultaneous flow event.

MUR adjustment is applied to measured consumption by Thames Water.

16 FMZ level calibration

16.1 Introduction

The Thames Water region is split into 6 WRZs and about 240 flow monitoring zones (FMZs).

The FMZ calibration provides a rigorous test of the model's performance by applying the same model used at WRZ level to a smaller zone where we expect larger fluctuations in explanatory variables and the consistency of behaviour we get from averaging is lower. This provides a good test of resilience to outliers as well as repeating the model over a large number of different examples.

In all but the WRZ tests (which has only six zones) we have relied on data from the same source the model was derived from, but here the model is performing on the entire population. The WRZ level calibration was used to develop the DMC, so cannot be seen as an independent test of the whole model performance.

This calibration will also allow us to analyse the performance of the model in conjunction with various development stages of the DMC in an independent test which extends the baseline of meter penetration levels.

We take the full model and apply it to each FMZ, computing the residuals for the various model outputs. This gives us a much lower level test of the performance of the model than does the uncertainty calculation (in which the precision of the forecast can in some respects be considered a test), while still on a level which includes averaging-out effects, in contrast to the application of the model to single households. This is an advantage as we are emphatically not trying to predict how individual households behave, we are trying to model a mean behaviour. In effect we can consider this a test of the scaling properties.

Another advantage of the FMZ calibration is that we can look at whether there are geographical/spatial factors in consumption patterns. Specifically we may look for geographic factors in clusters of "outlier" behaviour.

16.2 Method

To detail the implementation a little further we discuss how this variation of the model was built.

The FMZ-level model was built on the data for 237 FMZ sites and processed by the model, as well as trials of alternate parameters (for example London and Thames Valley coefficients applied to all zones for comparison, or a comparison of methods for modelling metered coefficient) implemented in the R programming language. The data provided by the client was combined with local authority data, and a number of the inputs to the model had to be inferred, which must be borne in mind when looking at the results.

The following approximations were necessary to perform the inferences:

- That the local authorities overlapping FMZs are spatially homogeneous;
- The ratio between metered and total occupancy remains fixed across scales;
- Mean rateable value (RV) is the same across each zone (this is almost certainly not a good assumption, but it is sufficient to assume that the variation in mean between RV from FMZ within the same WRZ does not have a significant impact on the forecast).

We also had to make certain choices in the implementation to ensure the necessary elements were properly tested, principally in fixing the calibration coefficient by WRZ, over the obviously flawed choice of calibrating by FMZ.

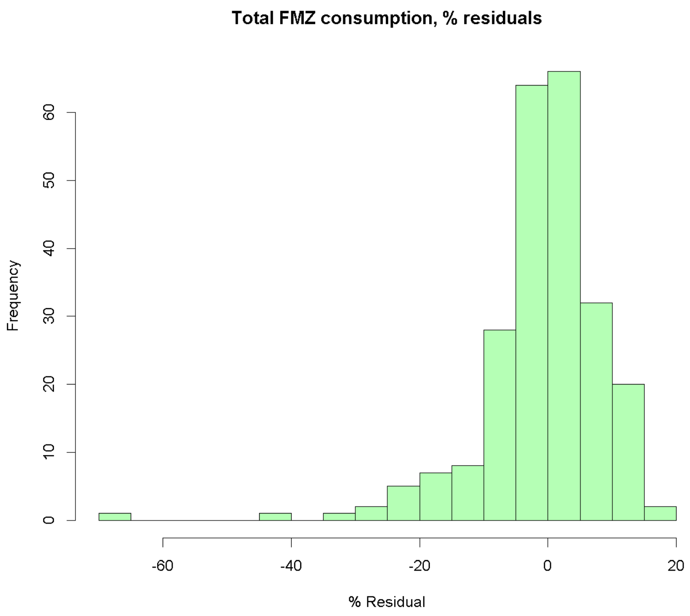
The following variations were used to test the model:

- As the final Excel model with dynamic metered coefficient and static intercept;
- As above but with double-dynamic metered coefficient;
- Coefficient testing (both zones run on London and Thames Valley coefficients); and
- Metered coefficients set to zero.

16.3 Results

Figure 32 presents a histogram of residuals on total modelled consumption at each FMZ as a percentage of total consumption (note that it presents total consumption over PHC, but in fact they are exactly equivalent in this case).

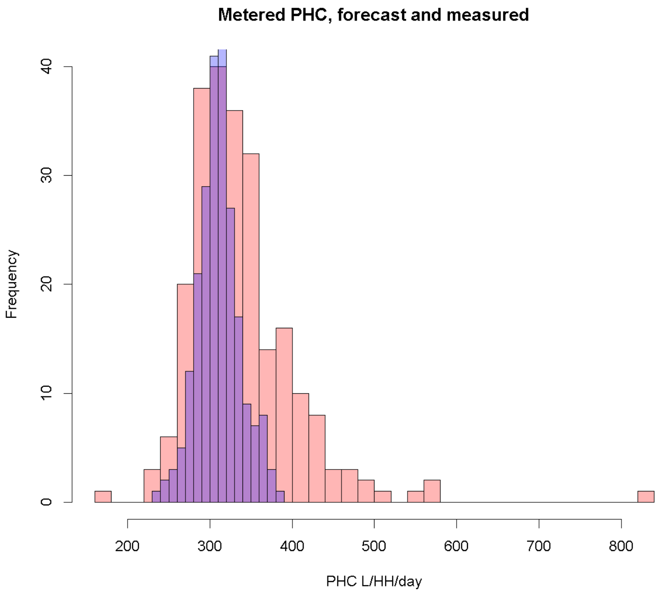
Figure 32 FMZ Modelling Residuals



The peak of the distribution sits around zero (the mean sits at -1.0%, but when we bear in mind the outliers being skewed negatively and use the median this shifts our central estimate to 0.014%). The distribution of results also suggests that the model performs well, if we put the long left tail to one side for a moment, as there seems to be a significant cluster around zero.

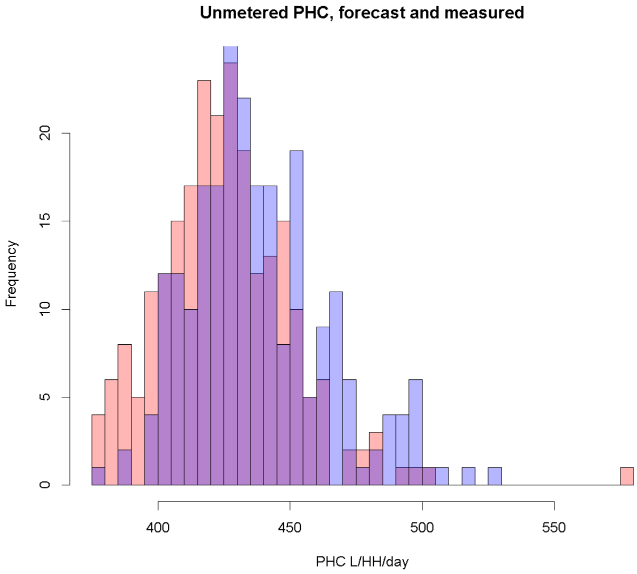
Focusing on the tail specifically the following plot sheds a little light. Figure 33 and Figure 34 show forecast figures in blue and the measured figures in red.

Figure 33 FMZ Metered Model Measured and Predicted PHC Distributions



What is striking about this plot is that while the tails of the distribution are not particularly closely modelled, the forecast sits around the mean and is very much reminiscent of the distribution of a sample mean estimate. This really shows what the model is doing; when modelling what is essentially a population mean (PHC) we find that outlier behaviour is not very well modelled. This is as expected because at the scales we model (WRZ), spatial averaging has a significant effect. This is a technical insight, the real point being that on the level at which the data is modelled it appears to hold up to scrutiny in forecasting the mean.

Figure 34 FMZ Unmetered Model Measured and Predicted PHC Distributions

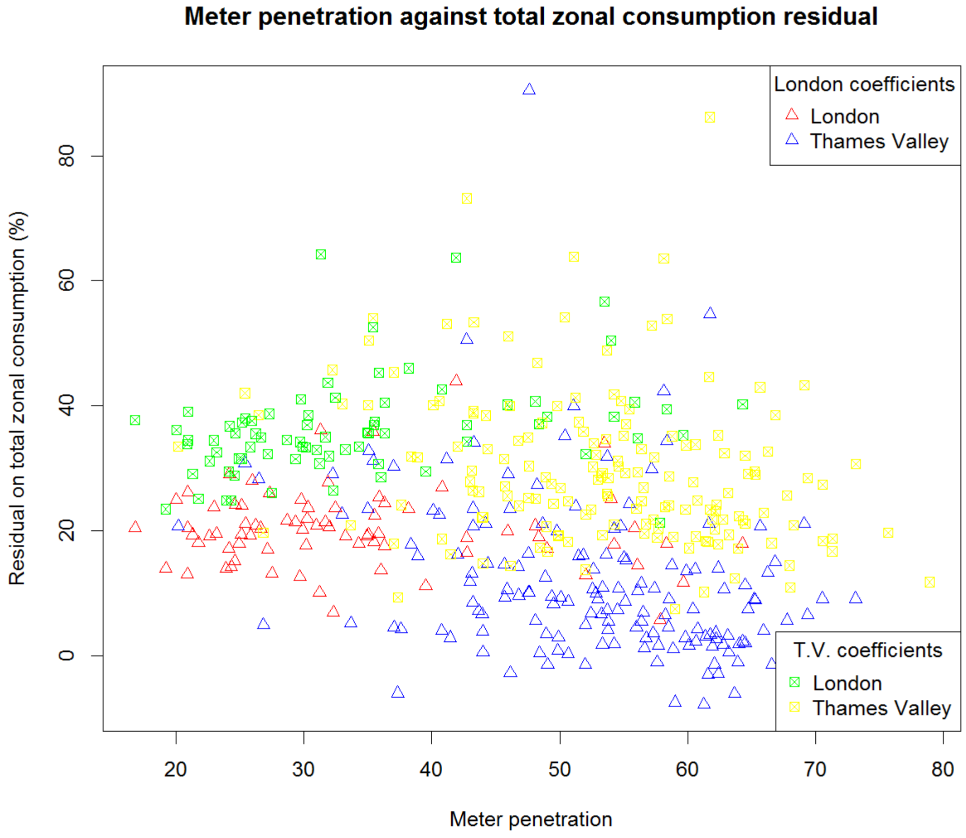


The unmeasured model expresses the dynamic range of consumptions very well, and this is likely because the base model was derived from the unmeasured population.

A more specific element that was tested in the FMZ-level analysis was the performance of the particular models, for the London area and for Thames Valley (note that by model we specifically mean the regression coefficients). This testing is done simply by changing the set of coefficients used in the calculation. We note that no calibration is done as the calibrations at WRZ level are performed to rescale the sets of coefficients, we therefore cannot simply test the coefficients by accuracy, we must test the forecast by its precision (note that this is not a litmus test, more of a gauge of performance). In this vein the following analysis explores the spread of the residuals.

Figure 35 shows total zonal consumption for the different sets of model coefficients.

Figure 35 Test of Thames Valley and London Model Commutability

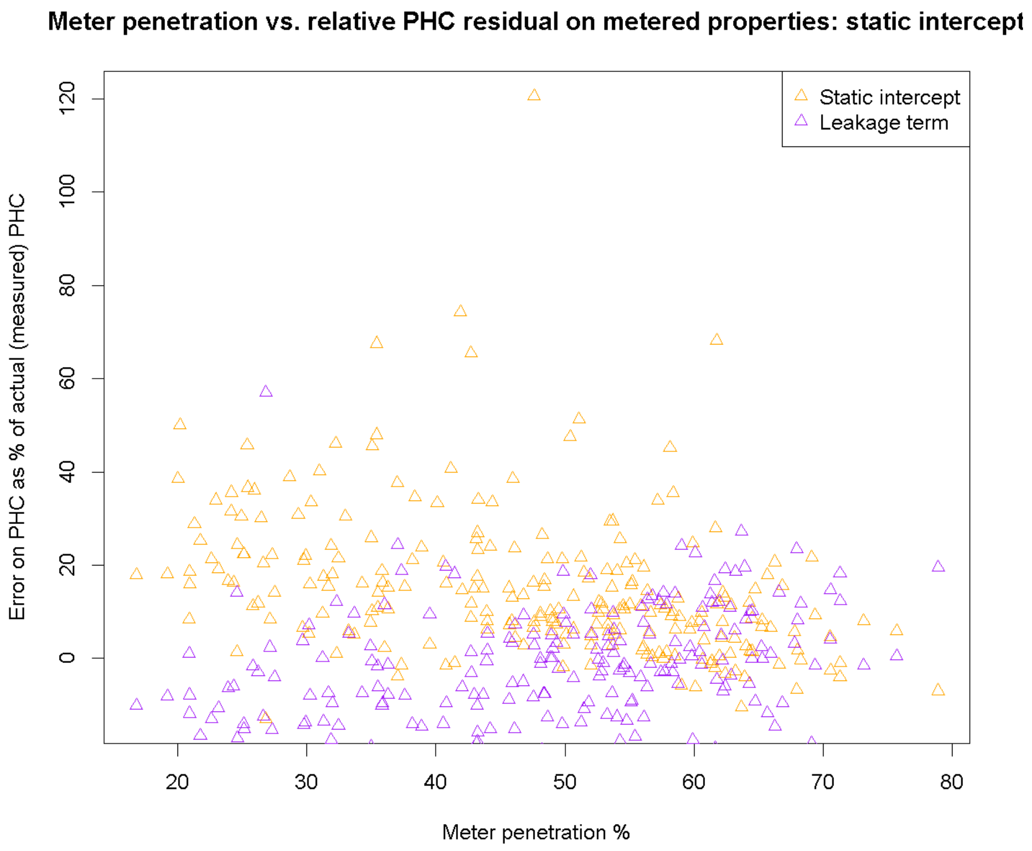


Though there isn't a stark difference in terms of the spread of the two models, analysis shows that in fact the London coefficients give a smaller standard deviation of percentage residuals, which suggests that they give a more precise forecast.

16.3.1 DMC Analysis

This lower level resolution allows us to explore the effects of different dynamic metered coefficient models. We particularly look at the effects of changing the linear dynamic metered coefficient with constant intercept to the same model using the new 'losses term' replacement for the intercept. Below is the zonal meter penetration against the residual on PHC for all zones in the model with the static intercept and the updated leakage term replacement

Figure 36 Performance Impact of Exponential Losses term on DMC

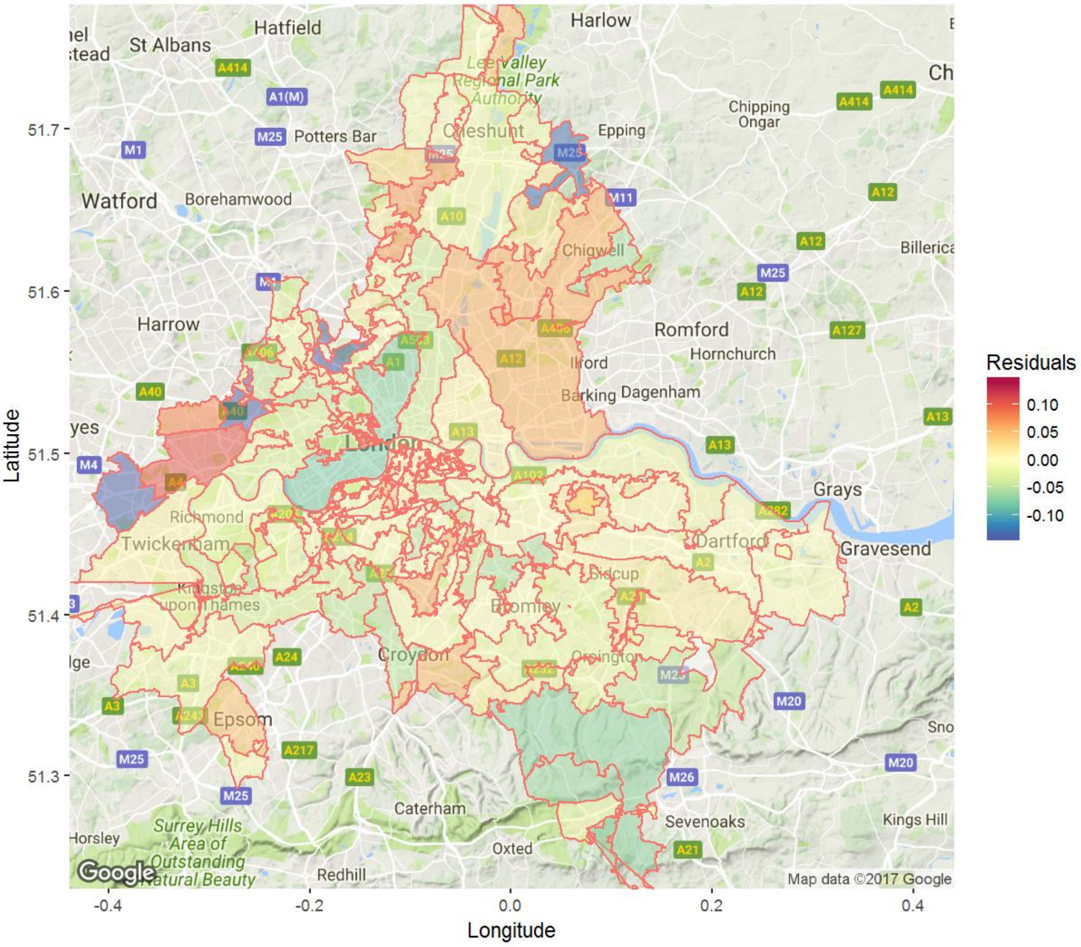


The losses term provides a very visible improvement in the model's performance, up to the higher end of meter penetration. At higher penetration the difference is less marked (and the new model's residuals seem to go up), but the plot suggests that further analysis of the new model could provide significant improvement.

16.3.2 Spatial Testing

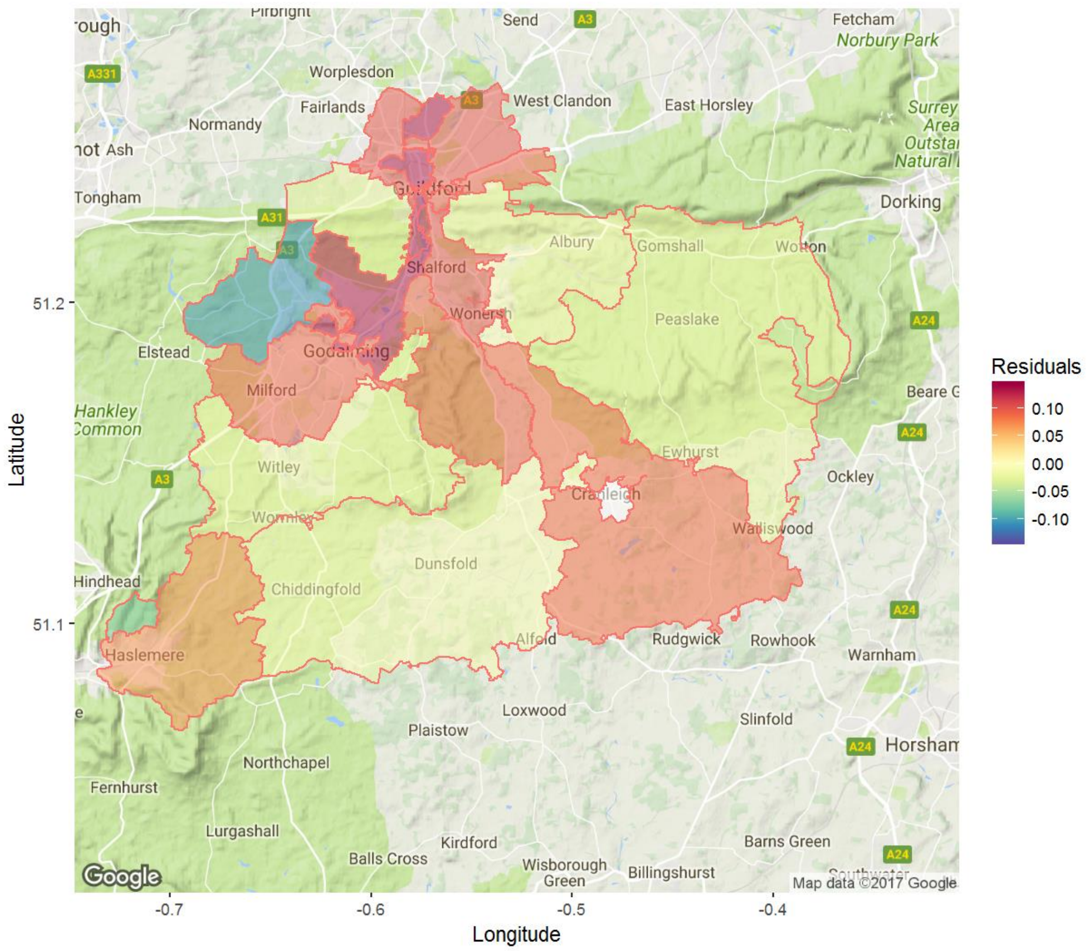
The final significant area to cover is the spatial testing. The analysis was performed in the R language to produce "heatmap" style plots showing the FMZs coloured by their over or under-estimation, quantified by the percentage total consumption residuals. The plots are split into three zones to make the presentation as clear as possible.

Figure 37 Heat map of FMZ model performance London Area



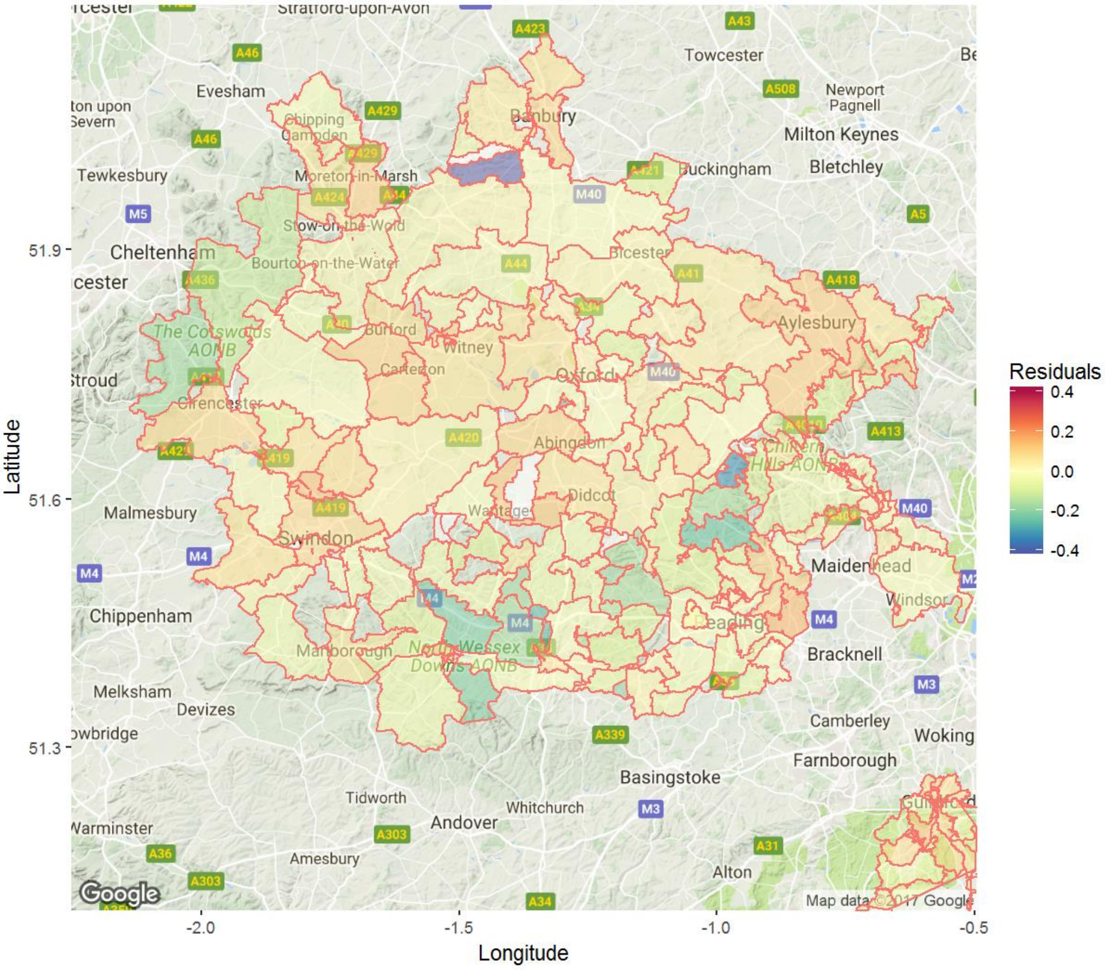
In the above plot we see the London area. The colour scale shows that the majority of errors are below 10%, lending veracity to the earlier claim that the London area coefficients perform very well. There is a small region in the west-north-west area of the map in which there is under and over-predicting zones next to one another. We do not as yet have sufficient data to determine the exact cause. However we observe that the region would appear to be predicted well if added together; this may be related to the inference of data.

Figure 38 Heat map of FMZ Model Performance Guildford Area



Next, the Guildford area has a similarly small range, suggesting that here too the model performs well, with perhaps a slight leaning toward over prediction.

Figure 39 Heat map of FMZ Model Performance Thames Valley Area



Finally, for the plot of the remaining zones. shows the scale choice is significantly broader, indicating a significantly larger residual than in the preceding figures(note for context that the Guildford zones are visible in the South East corner of the map).

A particular point of note on this plot is that there are two “streaks” of under-prediction, one along the Cotswolds (North West corner) and another along the Chiltern hills to the North Wessex downs. On closer investigation we see that a number of these outliers are in the tail of the metered PHC histogram (measured data) and that the metered PHC is significantly higher than the unmetered. This striking effect warrants further investigation but is presented as it appears both significant and persistent. It is thought that this may be related in part to the lack of FMZ resolution rateable values.

16.4 Summary

As a high level summary, the FMZ calibration indicates that the model performs well on not only lower level but also previously unseen data. Given the constraints on accuracy by the inference of data and the assumptions made, as well as the change in scale (given an expect greater fluctuations in forecast at lower level) the results appear to be in line with what the uncertainty calculation (see ‘Uncertainty’) would lead us to expect.

More broadly the FMZ calibration has proved a useful tool in optimising the model features, and the insight we gain into modelling choices provides a wide range of potential investigations to further tweak the model.

17 Climate change

UKWIR report 13/CL/04/12 “Impact of Climate Change on Water Demand” Provides a mean estimated linear trend for the effect of climate change on demand between 2012 and 2040 for England, for normal dry and critical periods.

The UKWIR estimate was drawn from a number of studies that modelled consumption responses to weather metrics. The assumption was made that the behavioural response to short term changes in weather would translate to long term changes.

Climate impact is outside the scope of this model, but provides the necessary zonal consumptions to which the proportionate climate impact can be applied.

18 Trends, scenarios and uncertainty

18.1 Trends

18.1.1 Calculated and observed trends

The household consumption model residuals from previous years produce a significant trend in time series, indicating that some of the projected change in consumption is not accounted for by dynamic time series parameters within the model such as occupancy rates and meter penetration.

This observed un-modelled trend is thought to be driven by technological and behavioural changes. A trend would also be produced when occupancy figures are incorrect or not updated.

In our assessment of future trends, we consider both observed, (which is a real deviation from our modelled output), and calculated trends, which are free from back cast model parameter error.

Some consumption trends are relatively well understood; a principal significant trend is a continued reduction in toilet flush volumes. The endpoint of the current trend is predictable, as more toilets are replaced that meet water regulations. Even this is not simple, and there has been recent evidence that the newer designs of cistern are more prone to silent leakage, reducing the expected water savings. If however this leakage is near meter stall sped, these losses will not register as consumption.

The cistern trend alone would present a considerable analysis problem if it were to be fully modelled.

It is also likely that showering and bathing water use is increasing. The water efficiency of white goods is improving, and the net trend of these various components is observed over the last ten years in the model residuals.

Projecting a linear trend ad infinitum is a nonsense, and will result in negative consumptions, but it is possible to draw upon observed data from micro-component studies carried out in 2002-04 and 2015-16 and project micro components to likely end and midpoints to derive a number of trend scenarios which can be superimposed over model outputs.

This analysis has combined observed micro component trends with calculated endpoint scenarios to devise a range of possible trends. Of these, the upper and lower trends are used to bracket a most likely forecast.

18.1.2 Trend scenarios from DWUS data

Separate trends were derived for both London and Thames valley data sets. The trend was observed over a ten year period and is derived from model residuals meaning that it is un-modelled. The trend was generally downward and stronger for winter; the year round trend was applied to the initial forecast period to 2045 and also for a ten year initial period.

It is possible that the strong Thames Valley and weak London trends could switch because evidence showed more efficient appliances are installed in Thames Valley, thus leaving a greater potentiality for consumption reduction in London area.

18.1.3 Micro component based trend scenarios to year 2100

Three of the scenarios are driven by observed trends carried through to 2045. Thereafter a year 2100 endpoint scenario is devised by bundling consistent outcomes for the various components and a direct trend is then drawn to 2045.

Devising endpoints for an 85 year trend is going to produce some fairly extreme scenarios: to put it in perspective, the difference between 1930 and 2015.

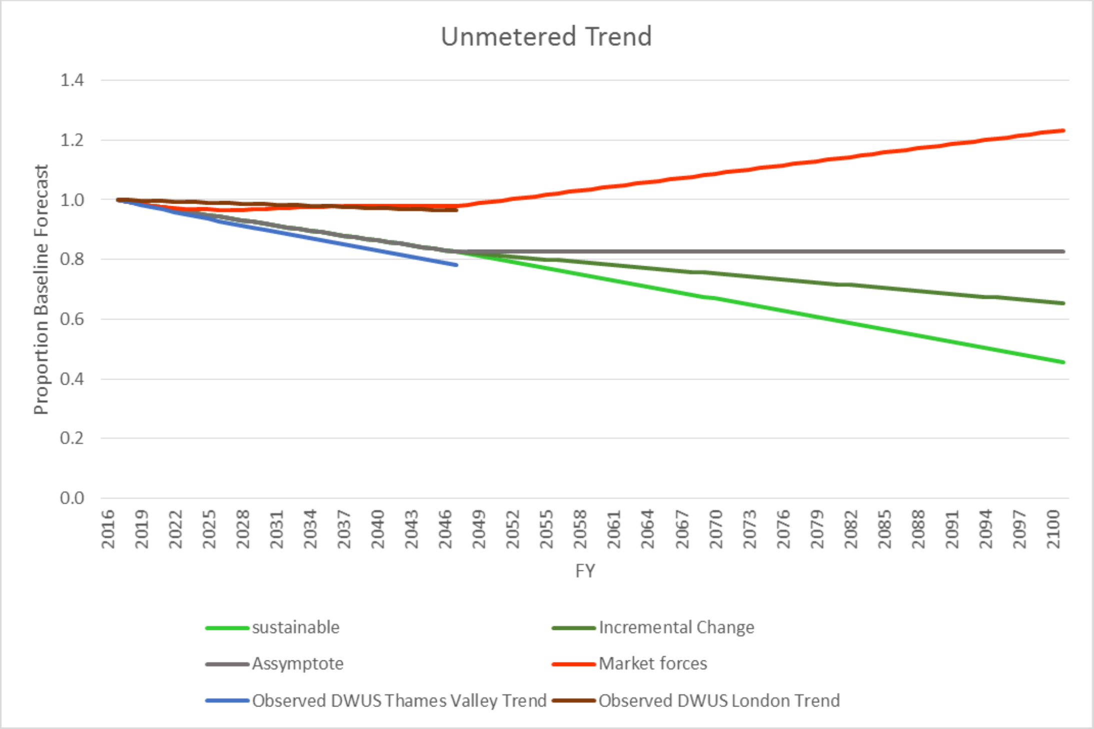
There is a spectrum from “market forces” trend where personal finances are prioritised over environmental considerations to sustainable development where environmental considerations are paramount.

It is considered that the central tendency, i.e. trends 2 and 3 are more likely outcomes than extreme trends 1 and 4. This can be quantified by Monte Carlo analysis of individual micro components to produce a continuum of trend bundling.

1. Sustainable Development - In this most extreme efficiency scenario, we have assumed that water saving is driven by both technological advancements and attitudinal changes. Sophisticated filtration technology would allow recirculation of shower water saving both energy and water. Waste water and washing functions are fulfilled by greywater recycling, aided by hydrophobic frictionless surfaces. Bathing is pretty much obsolete
2. Incremental Change - This scenario assumes that the current paradigm of regulatory driven incremental technological efficiencies will continue past 2045 and arrive at an endpoint that is conceivable with existing technologies but currently not economically viable.
3. **Asymptote** - In this neutral scenario, we make the fewest possible assumptions: Artesia’s measured micro component trend projection continues to 2045 before levelling off.
4. **Market Forces -** This scenario assumes that the projected trend in micro components does not continue beyond 2022. This would require a situation such as ‘hard Brexit’ where UK building regulations may be decoupled from current standards and the logical decline in flush volumes is curtailed. The observed upward trend in showering continues to increase.

The measured and unmeasured micro component trends are near identical; the graph below shows unmeasured trends because DWUS trends are derived from unmeasured properties.

Figure 40 Calculated and observed residual trends



- It is evident that there is broad agreement between the Thames Valley and most probable micro component trends, and the observed DWUS trends bracket the maximum likelihood trend.

- Micro component analysis provides evidence that the observed model residual trends may reasonably be extended beyond the observed ten year window.

Because of the pitfalls of the different approaches to trend, it is recommended that the central trend is used, which is a mean of contributing alternative scenarios. The most extreme trends can be regarded as minimum and maximum estimates.

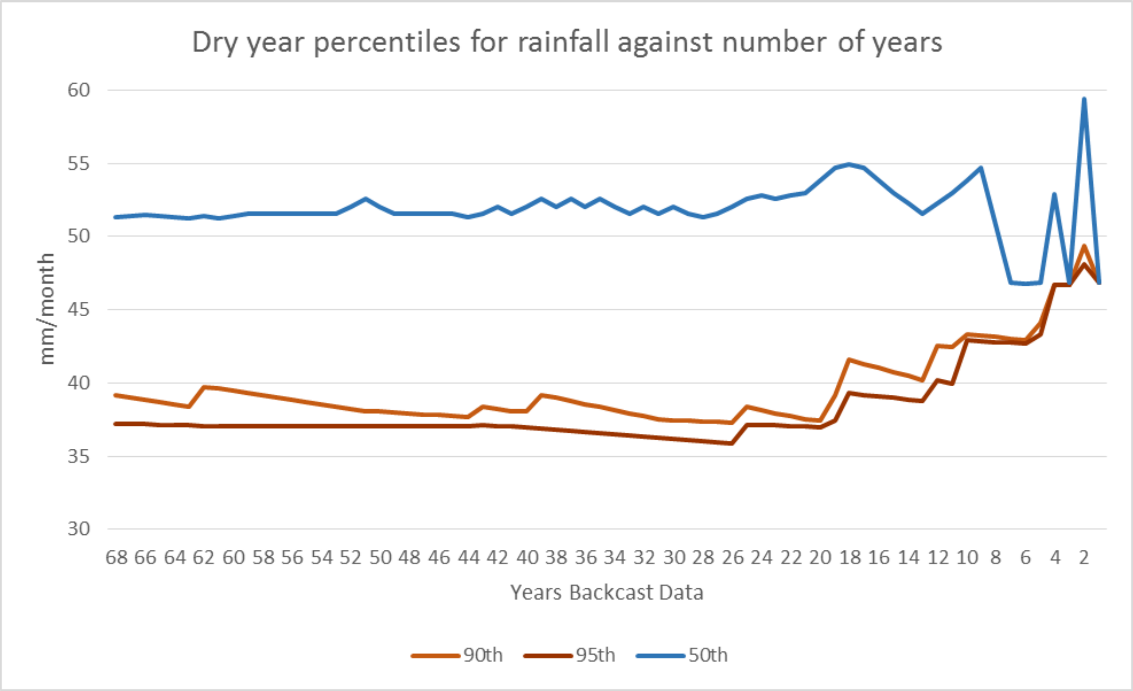
It is then recommended that as a central forecast, a mean of the observed and calculated trends are used, with the observed trend influencing the result for the first 20 years of the forecast where it aligns with calculated trends.

The contribution of trends to forecasting and uncertainty will be further considered in section 20.

19 Consumption uplifts for normal, dry year and critical period

The model baseline is FY 2015/16. It has been noted that model residuals are small for this year (14.2, Base year normalisation), and in addition the rainfall percentile for base year is 52mm/month almost exactly the historic 50th percentile as it becomes stable with increasing sample size. The difficulty here is establishing a statistically valid sample period in years against underlying trend. Figure 41 shows this relationship. For these reasons Base Year 2015/16 can be considered a normal year.

Figure 41 Stabilisation of dry year percentiles in cumulative historic data



20 Uncertainty

20.1 Introduction

The uncertainty in the model is important to give context to the forecast figures. It assigns a level of confidence to the figures we expect to see and allows us to talk more broadly about what we expect than simple scenario testing. The uncertainty model is implemented both for scenario testing and for full forecasts. For the former we offer a more refined set of scenarios in which we have a distribution of forecasts given a scenario, and for the latter we have a distribution of scenarios which are applied to give a total prediction given all uncertainties.

It may seem at first glance that the purpose of the error calculation is to gain an insight into how well the model works-indeed this is a very useful feature of the calculation-but its primary function is to predict an interval of figures at a reasonable level of certainty.

We also mention the important caveat that the uncertainty estimate quantifies what our model suggests is likely to happen assuming data errors are from random sources; it does does not make assumptions on the choice of model (though a good model choice should-in principle-reduce the assessed uncertainty, and vice versa).

So in summary, this calculation provides a spread of forecasts with assigned levels of confidence.

20.2 Method

The following parts of the calculation are factored in to the uncertainty, as they are taken to have a significant impact on the model uncertainty:

- Population and property forecast uncertainty
- Model coefficients
- Demographic/property-type distributions
- Base year figures (e.g. occupancy)

The population and property figures are presumed to be the main driver of uncertainty on zonal consumption, as variations in PHC (and PCC) tend to be have a small impact compared to how much total consumption varies across population and property forecasts. In the data we see that the contribution of these forecasts to the total variance grows more significant over time (at the end of the forecasting period the standard deviation of the population and property only model is over 50% that of the full model).

This calculation uses the UKWIR method for populaton and property uncertainty (report ref: 15/WR/02/8) and takes the guideline distribution around the forecast figure. As the distribution around the figure is normal by assumption, we treat the forecast figure as the mean and apply the RMS error figure provided in the guideline as the standard deviation. This then fully specifies the distributions for each.

We also include a correlation between the population and properties; heuristically growth in properties enables and is encouraged by growth in the population (and vice versa). The correlation coefficient for the Thames Water area was computed and applied, treating population and properties as a two-variable distribution.

The remainder of the uncertainties are

- Model coefficients: treated as normally distributed with the standard deviation taken from the standard error of the coefficients

- Demographic/property-type splits: varied using same distributions as on population and properties, but re-normalised each time-step to avoid further variation of population
- On figures input for base year only (e.g. occupancy): assume the same error as the base year forecast by way of a measurement uncertainty

This fully specifies input uncertainties, the final stage being to combine the uncertainties into a forecast distribution. As direct calculation was impractical, a numerical approach was necessary.

20.3 Monte Carlo analysis

The calculation was done by Monte-Carlo simulation which has three significant advantages:

- Monte Carlo analysis gives a robust approximation to a calculated figure with minimal computational complexity.

- Outputs have a straightforward interpretation, in that the iterations of the calculation can be taken to correspond to a chosen scenario, with specified probability of some input values.

- The calculation returns a distribution of possible outcomes, allowing us to see the shape of the distribution without recourse to more complex means.

Monte Carlo Analysis is implemented in the Excel model using the add-in @RISK 7.5 which has the added feature of displaying the distributions of consumption figures, along with the percentiles and standard output statistics, though a statistical summary page is also included in the model.

The micro-component trend scenarios are applied as part of the Monte-Carlo, using a PERT distribution with the upper and lower trend scenarios as the 90th and 10th percentiles, the peak being defined by the mean of the two central scenarios.

We have also included a variation on the main model calculation for testing the effects of specifically population and property variations on the demand forecast.

20.4 Analysis results

The following outputs show the central estimate for zonal consumptions, bounded by 5th and 95th percentile probabilities. In interpreting these outputs it is important to understand that this is an envelope for trend- it is not within these probability percentiles that consumption would move from the upper to the lower bound in a single timestep. As points are fixed in the orderly transition from future to present, the range future possibilites at any given year diminishes.

Figure 42 Thames Valley Monte Carlo Forecast

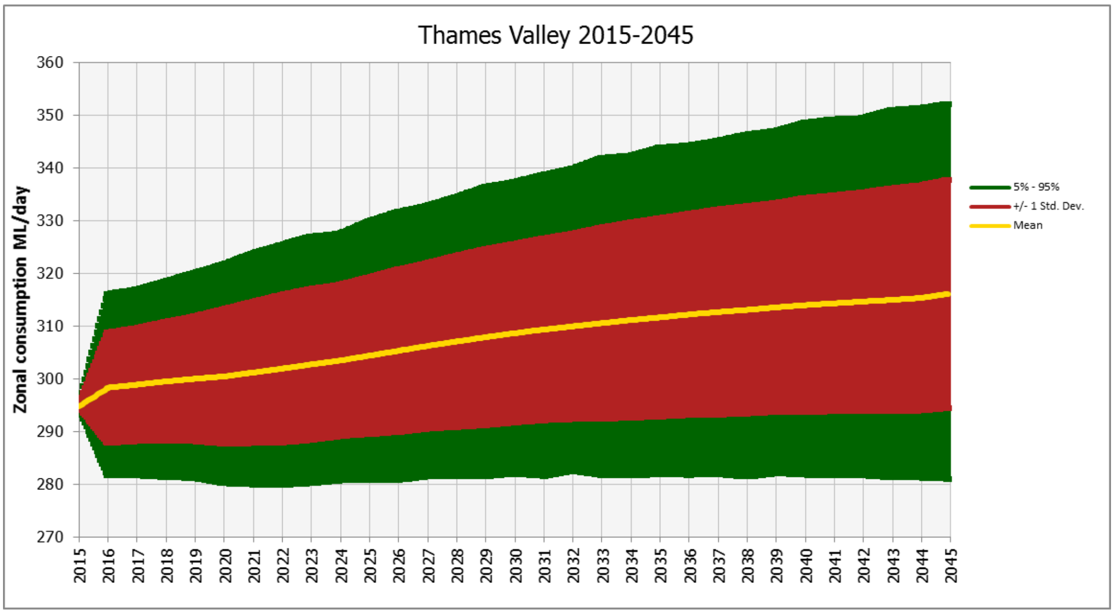
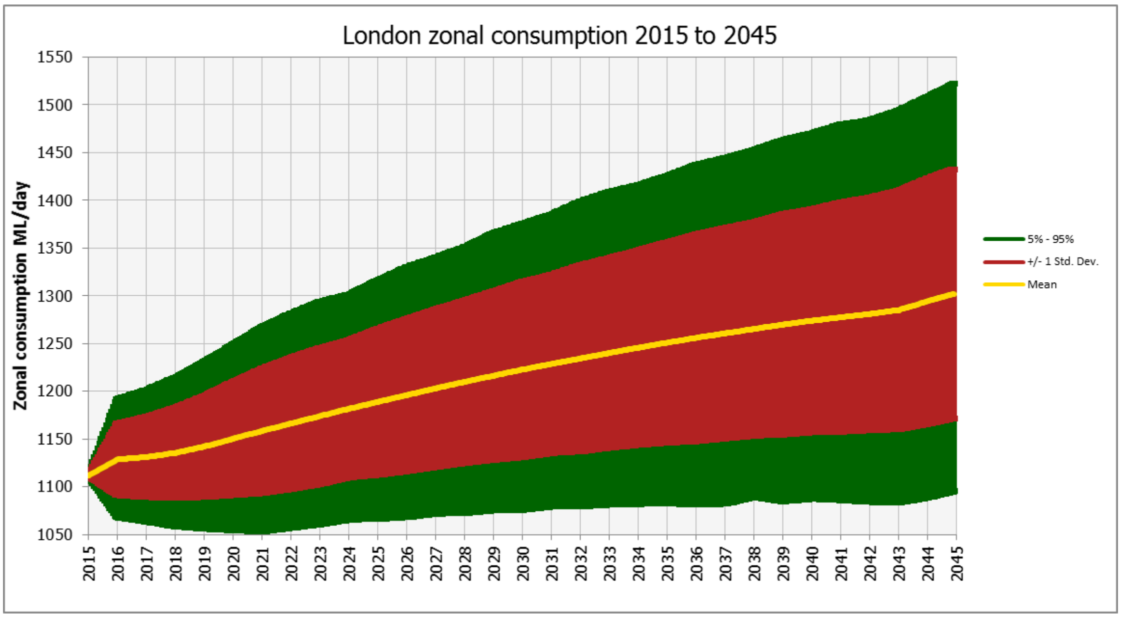


Figure 43 London Monte Carlo Forecast



The first point to notice is that the base year (2015) has zero error. This is because the base year is modelled and then calibrated to a reported consumption figure to which we do not assign an uncertainty.

The second observation is that the breadth of the distribution grows as a function of time and commensurate increase in uncertainty.

The results, particularly in light of the population and property sensitivity testing, show that the uncertainty on the regression coefficients makes a significant difference in the earlier years of the forecast, until eventually the population and property forecast errors become dominant.

This is corroborated by the simulations run on only the population and property variations, which are shown below.

Figure 44 Thames Valley Population Uncertainty Based Forecast

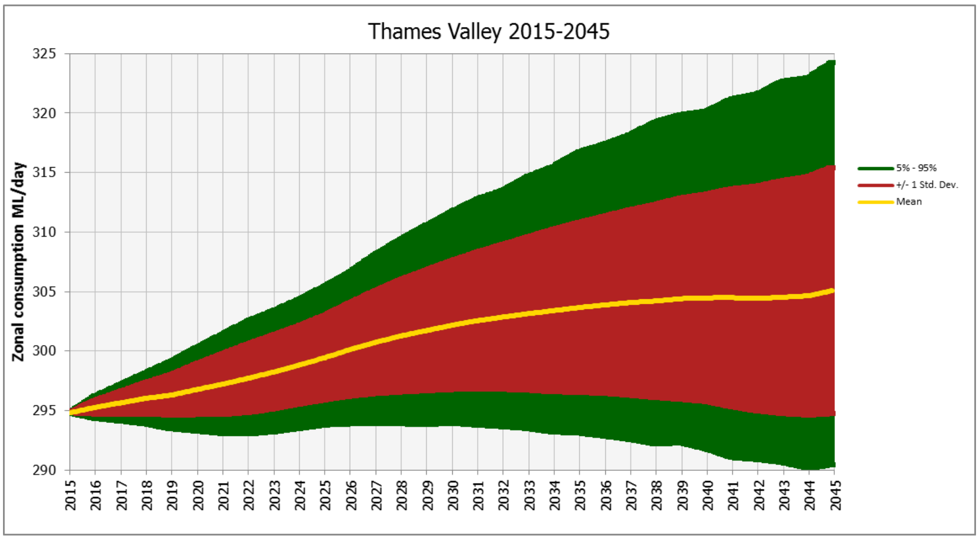
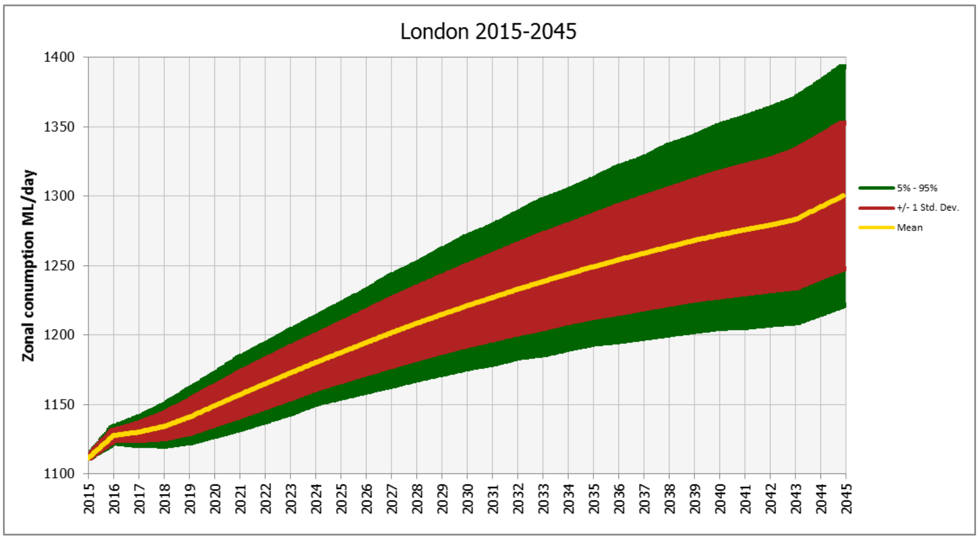


Figure 45 Thames Valley Population Uncertainty Based Forecast



These plots are similar but essentially just isolate the component that grows with time.

20.5 Summary

To summarise, the uncertainty for the MLR model completes the forecast in the sense that it gives an assessment of all the possibilities the data suggests. The plots suggest that in the earlier stages the uncertainty is principally from the model error (though there is a strong argument that in the very early stages this would be a conservative estimate), but over time this is dominated by the accuracy of the population and property forecast.

21 Conclusions and recommendations

Multiple linear regression (MLR) modelling has been successful in producing a robust model that integrates the impacts of drivers such as occupancy, property type, socio-demographics and meter penetration in a dynamic model which can be used to test sensitivities around individual parameters.

In addition to the functional advantages of an MLR model, model error can be quantified and model performance tested throughout the build process. Iterative model building can hone the performance of the model at individual and aggregate level.

Within this report we have validated the MLR models using three different approaches. Firstly the model is constructed using standard statistical methods from which the uncertainty can be quantified. Secondly the model has been validated temporally, by applying the model to historic data and forecasting forwards to the current year and comparing with reported figures. Thirdly, the model has been validated spatially, by applying the model to about 240 sub-zones across the Thames Water region and comparing with reported data. This is a level of validation that could not be carried out with previous micro-component based models.

There is certainly potential for the MLR consumption modelling tool to be developed and bridge the gap between strategic and operational planning.

There are a number of modelling and forecasting areas with high uncertainty that can be addressed in order to tighten the forecast envelope. Many of these are dynamic, in the sense that they may change continually over time and so require time series data which is gathered with uniformity both to understand them and forecast them.

As it stands, uniformity of data gathering is important for time series, so continuing the general methodology for demographic and occupancy calculations is important. Improving data gathering can create a false time series trend. Estimated occupancy can be correlated against questionnaire answers if and when a smart monitor is created so that the impact of data collection methodology may be understood.

Artesia recommend more targeted higher quality data gathering and updating using forecastable parameters of proven predictive value in preference to the more general DWUS demographic dataset. Accurate estimates of occupancy of the various meter strata is of key importance in trend analysis of back cast model residuals, which is a cornerstone of forecasting trend.

Suggested studies for improving future forecasts:

1. Smart Monitor

As metering is rolled out there will be an opportunity to quantify measured consumption and losses. The change in occupancy, consumption and losses across the meter adoption process will be of the greatest importance in forecasting absolute savings due to metering and requires a high temporal resolution in occupancy data.

Measured data can be as useful as DWUS data if the modelling parameters are attached to the properties; therefore it is important that these parameters are collected and associated with consumption from measured properties.

A smart monitor must likewise be used to quantify changes between dumb and smart metering; it is in effect a randomised trial when customers who have previously opted for a meter are subsequently randomly upgraded to a smart meter, and will give a pure assessment of the impact of smart metering on consumption.

It would be especially useful to estimate continuous flows before the smart meter was installed on these properties. This study however has one systematic error- the difference in performance of the two meters. This would need to be addressed in the study, ideally by leaving the old dumb meter in situ.

2. Losses Study

Smart meters will obviously help in determining the proportion of consumption that is continuous, and whether the meter is positioned at the boundary or indoors will determine whether calculated losses are plumbing losses or all losses.

On large zones, the predictive power of the consumption model could well be limited by the accuracy of reporting, an element of which is the losses or leakage model, so the accuracy of consumption modelling is predicated on losses/leakage modelling.

3. Longitudinal consumption studies

Artesia have observed in a number of data sets that the initial consumption reduction in optant properties is not necessarily sustained. Further it is not known how this relates to occupancy and change of ownership. Furthermore, it is not known whether an occupant who moves into a measured property from an unmeasured property will show a reduction in consumption.

Most studies are discontinuous in that the data series ends with a change of occupant; indeed the ability to track customers as they move between properties, and properties as customers move in and out would shed light on the interaction between metering, optancy, occupancy and property.

Longitudinal consumption studies, i.e. those studies which track properties and occupants over time, measure their consumption and relate this to policies, characteristics, lifestyles, etc., are extremely important for improving future forecasts. These should be well managed and maintained, and should follow both unmeasured, and different cohorts of measured households.

4. Rateable value per FMZ

One issue with the FMZ calibration is that the mean RV was not available by FMZ. Although the contribution to the MLR model output from the RV coefficient is relatively small, it is also the only feature providing an economic proxy, therefore dropping variations between FMZs gives the model less data on potentially very significant behavioural drivers.

22 Appendix

It can be observed that there is some winter data bias in presence of consumption data with associated valid survey by month (below), and slightly more results are also rejected due to survey invalidity in winter (Figure 47)

Figure 46 Monthly Data points available

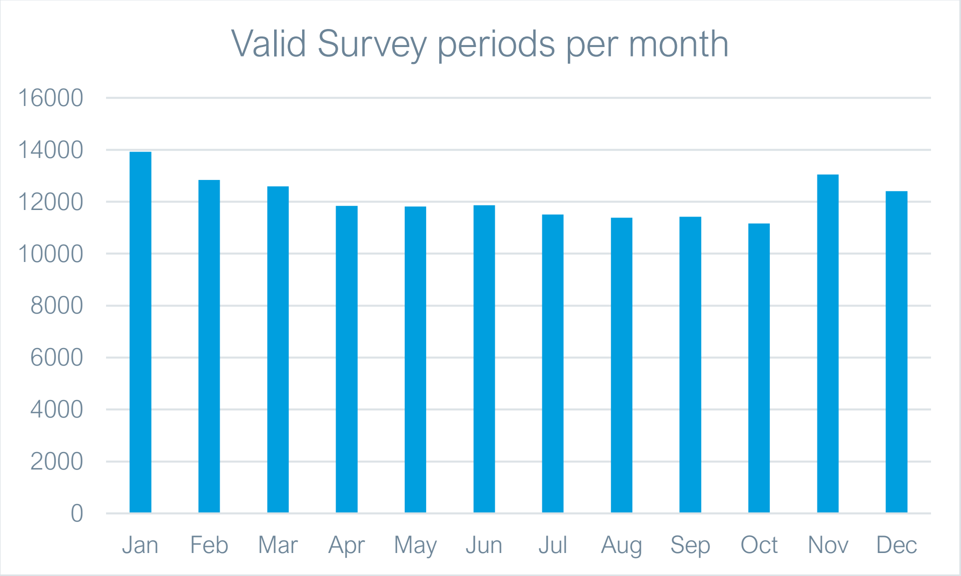
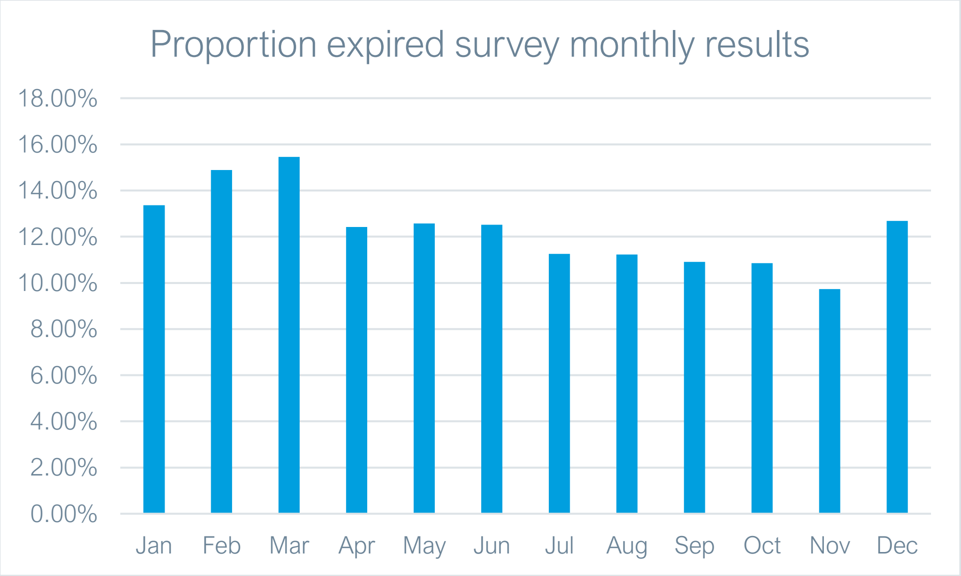


Figure 47 Invalid Surveys associated with flow data by month



Despite this observed bias, the model build is not affected, as is shown by the following test:

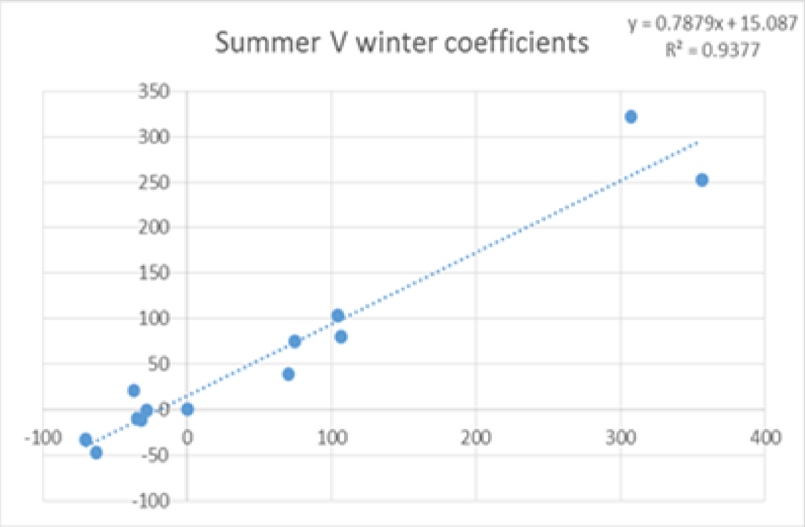
22.1 Testing seasonal variation

One approach to modelling consumption is to model winter and summer use separately. This has a valid basis in that seasonality response may not be uniform across all strata. Separate modelling would reconcile the different responses according to the proportion of these strata by zone. One example would be the expected response to warm dry periods in flats compared with the response from detached houses with gardens.

To test the hypothesis that the final model has captured the seasonal variation adequately, summer and winter trainer data was used to derive independent seasonal models, each comprising of six months' data. The combination of these models will be compared with the final model to see how the coefficients vary, and will provide a different kind of validation to the household model.

The separately derived coefficients for the summer and winter models were distinctly different, and are shown plotted against each one another in Figure 48.

Figure 48 Coefficient deviation of summer and winter models



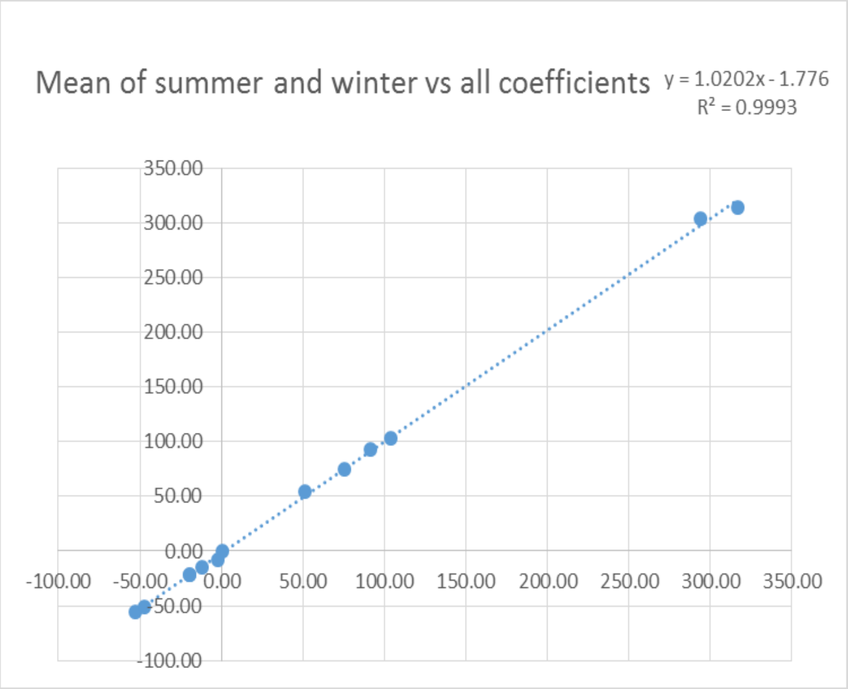
Since only annual consumptions are required for the forecast response variable, the annual model return would be predicated on the mean of the two sets of coefficients (each representing six months).

When the seasonal coefficients are averaged, and compared against the coefficients derived from the original unified data in the final model, a near exact match was achieved (Figure 49)

This is what we'd like to see, as it indicates the internal consistency of the methodology and the independence of the variables used in model derivation. Although this is an unconventional test of robustness, it is a good way to highlight any problems in the model, and would justify further testing.

This method also indicates that there is no advantage to be gained in annual prediction by building and combining separate winter and summer models.

Figure 49 Coefficient deviation of combined summer and winter models vs all year model



Thames Water