Citation for published version:
Xenochristou, M, Hutton, C, Hofman, J & Kapelan, Z 2020, 'Water demand forecasting accuracy and influencing factors at different spatial scales using a Gradient Boosting Machine', *Water Resources Research*, vol. 56, e2019WR026304. https://doi.org/10.1029/2019WR026304

DOI:
10.1029/2019WR026304

Publication date:
2020

Document Version
Peer reviewed version

This is the peer reviewed version of the following article: Xenochristou, M., Hutton, C., Hofman, J., & Kapelan, Z. (2020). Water demand forecasting accuracy and influencing factors at different spatial scales using a Gradient Boosting Machine. Water Resources Research, 56, e2019WR026304., which has been published in final form at https://doi.org/10.1029/2019WR026304. This article may be used for non-commercial purposes in accordance with Wiley Terms and Conditions for Self-Archiving.

Water demand forecasting accuracy and influencing factors at different spatial scales using a Gradient Boosting Machine

M. Xenochristou[1], C. Hutton[2], J. Hofman[3], and Z. Kapelan[1, 4]

[1] Centre for Water Systems, University of Exeter, North Park Road, EX4 4QF Exeter, U.K.

[2] Wessex Water, Claverton Down Road, BA2 7WW Bath, U.K.

[3] Water Innovation and Research Centre, University of Bath, BA2 7AY Bath Avon, U.K.

[4] Delft University of Technology, Stevinweg 1, 2628CN Delft, Netherlands

Corresponding author: Maria Xenochristou (mx220@exeter.ac.uk)

Key Points:

- The Mean Absolute Percentage Error ~~increased~~ increases exponentially from 3.2% to 17% for a reduction in group size from 600 to 5 households.
- Past consumption data and household characteristics ~~were~~ are important predictors of consumption for smaller aggregations of properties.
- The weather influence on consumption only ~~became~~ becomes visible for larger aggregations of properties.

Abstract

Understanding, comparing, and accurately predicting water demand at different spatial scales is an important goal that will allow effective targeting of the appropriate operational and conservation efforts under an uncertain future. This study used data relating to water consumption available at the household level, as well as postcode locations, household characteristics, and weather data in order to identify the relationships between spatial scale, influencing factors, and forecasting accuracy. For this purpose, a Gradient Boosting Machine (GBM) was used to predict water demand 1-7 days into the future. The results obtained show an exponential decay in prediction accuracy from a Mean Absolute Percentage Error (MAPE) of 3.2% to 17%, for a reduction in group size from 600 to 5 households. Adding explanatory variables to the forecasting model achieved a reduction in MAPE of up to 20% for the peak days and smaller household groups (20-56 households), whereas for larger aggregations of properties (100-804 households), the range of improvement was much smaller (up to 1.2%). Results also showed that certain types of input variables (past consumption and household characteristics) become more important for smaller aggregations of properties whereas others (weather data) become less important.

Keywords: water demand forecasting, Gradient Boosting Machines, spatial scales; smart demand data; weather influence;

1 Introduction

The effectiveness of future efforts, technologies, and conservation strategies is heavily dependent on accurately predicting water demand at the appropriate scale. From emerging technologies (e.g. gray water recycling at the household level) to conservation campaigns (e.g. changing customer's attitudes) or even future investments (e.g. building new reservoirs), solutions are typically targeted at a certain level of spatial aggregation. Thus, accurately predicting demand at the appropriate scale is of the utmost importance for their success.

As part of the commitment to sustainably manage their water resources, water companies are required to reduce per capita consumption (PCC) and leakage, in order to reduce the impact they have on the environment (Ofwat, 2017). According to the Office for National Statistics, PCC in the UK is the 5th highest in the EU (Bailey, 2019), amounting to a total of 114 l/capita/day. Gaining a better understanding of the factors that influence water use at different spatial scales can assist with developing improved water demand management strategies and curbing demand. Leakage also remains at relatively high rates, as approximately 23% of the total inflow into the network is lost through leaks (Ulanicki *et al.*, 2009). Ofwat, one of the UK water industry's regulators, has challenged water companies to reduce this figure by 15% by 2025 (Ofwat, 2019).
Operators can choose to estimate leakage at different reporting levels, such as district meter areas (DMA), water resource zone levels or even an intermediate zone level within the distribution network (Ofwat, 2018). In order to do this, they need to be able to accurately forecast water demand at different levels within the network. Therefore, the forecasting accuracy that can be achieved at each level, as well as the factors that determine it need to be assessed. This will allow water companies to make informed decisions and their regulator to accurately assess their performance.

Field Code Changed

However, predicting water demand is not an easy task as there are many uncertainties involved in the process. The main challenges arise from the tight relationship between the human and natural systems in urban environments, where more than half of the population currently resides (House-Peters and Chang, 2011), as well as the many time- and space- dependent factors that can influence water consumption (Parker and Wilby, 2013). Furthermore, the maximum prediction accuracy that can be achieved as well as the most influential explanatory factors can vary greatly depending on the spatial scale. When aggregating large areas, the demand signal is fairly smooth since it averages out over a large number of water users. On the other hand, small-scale water use is likely to be associated with increased noise in the data, leading to a higher uncertainty and thus increased errors.

This study explores in detail and quantifies the relationship between spatial scale and demand
This study explores in detail and quantifies the relationship between spatial scale and demand
This study explores in detail and quantifies the relationship between spatial scale and demand
This study explores in detail and quantifies the relationship between spatial scale and demand
This study explores in detail and quantifies the relationship between spatial scale and demand

- What is the maximum demand forecasting accuracy that can be achieved at different spatial scales?
- What are the most important influencing factors at each spatial scale?

The current paper is organised as follows. The next section discusses the results and shortfalls of
The current paper is organised as follows. The next section discusses the results and shortfalls of
The current paper is organised as follows. The next section discusses the results and shortfalls of
The current paper is organised as follows. The next section discusses the results and shortfalls of
The current paper is organised as follows. The next section discusses the results and shortfalls of

2 Background

Several studies attempted to predict water consumption, using a great variety of data, models, methods, as well as explanatory variables (Prescott and Ulanicki, 2008; Herrera *et al.*, 2010; Adamowski *et al.*, 2012; Tiwari and Adamowski, 2013; Matos *et al.*, 2014; Romano and Kapelan, 2014; Hutton and Kapelan, 2015; Anele *et al.*, 2017; Brentan *et al.*, 2017; Zubaidi *et al.*, 2018; Xenochristou et al., 2020b). Some studies in the literature even accounted for the spatial variability of water demand (Balling at al., 2008; Lee et al., 2009; House-Peters et al., 2010; Polebitski and Palmer, 2010; House-Peters and Chang, 2011; Maheepala *et al.*, 2011; Rathnayaka *et al.*, 2017a; Chen and Boccelli, 2018). Lee et al. (2010) used space-time variation and projections on population density to forecast water demand for the city of Phoenix over a time-space dependent grid. Although integrating future estimates in the forecasting methodology improved the forecasting accuracy, Lee et al. (2010) argued that additional input factors (other than population density) could improve the forecasting accuracy. Rathnayaka *et al.* (2017a) introduced a model that predicts water end-uses for different types of households at multiple temporal and spatial scales. Although this approach made use of a variety of household, temporal, and weather characteristics as predictors, it did not deal with consumption at each scale as a separate problem. Instead, the total consumption was constructed by merely adding the individual end-uses of the households in each aggregation of properties. A study by Balling et al. (2008) investigated water consumption among census tracts and the effect that several weather variables have on it. Using a variety of explanatory variables, it concluded that census tracts' sensitivity to drought depends heavily on their socio-economic and land-use characteristics (particularly the presence of pools). However, results were only tested at the census tract scale. House-Peters et al. (2010) investigated the drivers of water demand in Hilsboro, Oregon and concluded that drought condition was not a good predictor of water use at the study area level, although it was for certain census blocks containing large, new, affluent, and well-educated households.

As it becomes apparent, although few studies implemented spatial variability in their forecasting models, there are certain limitations. One of the limits for comprehensive spatial analysis of water demand has been data availability at high spatial resolutions or in many cases the level of spatial aggregation of water consumption data not matching the scale of the explanatory variables. In order to overcome this problem, researchers often have to rely on interpolating or extrapolating data (Lee at al., 2010; House-Peters and Chang, 2011), i.e. estimating values for locations within the study area or outside the study area, respectively, which can be a challenging process (Lee at al., 2010). Even when data is available at the household level, it often lacks spatial coordinates (House-Peters and Chang, 2011), sometimes due to privacy concerns. Another main problem derived from the current literature is the lack of a systematic comparison of accuracy and influencing factors at various spatial scales. Since the variables that influence water consumption and the range of temporal and spatial scales can vary greatly at different settings and case studies, this comparison cannot be derived by merely comparing the results of different studies in the literature. To summarise, although a substantial increase in data availability, computational power, and new technologies over the recent years has contributed in developing spatially explicit demand forecasting models and identifying and quantifying relationships among a variety of weather, social, and water consumption data (House-Peters and Chang, 2011; Rathnayaka *et al.*, 2017; Xenochristou et al., 2020b), there is still the need to develop methodologies that incorporate this information at multiple spatial scales (House-Peters and Chang, 2011).

This study aims to address this gap by making use of a very rich dataset comprising of a variety of household characteristics, weather data, temporal characteristics, and past consumption. The aim is to use these data to identify and quantify the influence of the drivers of water demand at multiple spatial scales and determine how they contribute to the accuracy of demand forecasting models.

3 Data

3.1 Data Description

The consumption data comes from a region in the southwest of England and relates to 1,793 properties. These were monitored by the water company using smart meters at 15-30 minute intervals, over a period of almost three years (October 2014 to September 2017). The raw dataset was carefully cleaned in order to exclude incorrect and missing data, empty properties, and leakage. A detailed description of the cleaning process can be found in Xenochristou et al. (2020a).

The water company also collected data related to the households' characteristics and partial postcodes. Information regarding the garden size, occupancy rate, metering status, rateable value of the property, residents' socio-economic status (ACORN), and council tax band became available at the household level. The occupancy rate of the household refers to the number of people living in the property, whereas the metering status reflects if the property is billed based on their meter reading or not. In the UK, approximately half of the properties are unmetered (Xenochristou et al., 2020a) and their water bill is calculated based on an estimation, partly dependent on the property's rateable value. The higher the rateable value of the property, the higher the water bill (for unmetered properties). ACORN is a geodemographic segmentation of the UK's population in customer types, based on social factors and population behaviour (CACI Limited, 2014). According to the ACORN guide, customers are divided into groups A to Q, with groups A to E classified as affluent, F to J as comfortable, and K to Q as financially stretched. The council tax band reflects the council tax rate the property belongs to, based on its location. Council tax bands vary from A to H, from the lowest (A) to the highest (H) paying band. The garden size is the size in m^2 of the property's garden. Finally, postcodes in the UK are comprised of four parts, indicating the area, district, sector, and unit the house belongs to (Royal Mail, 2012). In this study, only the first two parts of the postcode, corresponding to the area and district, were available and used to group the properties.

Each one of the above six household characteristics (garden size, rateable value, occupancy rate, council tax band, rateable value, and ACORN group) divides the dataset into different categories, depending on the individual attributes of each household in the dataset. For example, depending on the characteristic 'garden size', the households are divided into three categories, 'large', 'medium', and 'small', reflecting the size of the garden of the corresponding household. The categories created for each household characteristic are presented in Table 1. Out of all six characteristics, two of them (garden size and metering status) were Table 1. Out of all six characteristics, two of them (garden size and metering status) were Table 1. Out of all six characteristics, two of them (garden size and metering status) were organised into categories by the water company, whereas the rest of them (rateable value, acorn group, occupancy rate, council tax band) were divided by the authors. The aim in forming these categories was to create groups that were large enough to be representative, while at the same time being distinct enough from the rest of the groups to offer a certain explanatory value. A z-statistic was used here to assess the similarity between the groups. For example, the similarity between the distributions of daily consumption values over the three years in the data between council tax bands A, B, and C was assessed using a z-statistic and was deemed similar enough to group them together into category A-C.

Furthermore, weather data on air temperature, soil temperature at 10 cm depth, humidity, sunshine duration, and rainfall became available by the UK's Meteorological Office (Met Office). These data were recorded at the hourly or daily scale over the same period (October 2014 to September 2017), from hundreds of weather stations across the study area, as part of the Met Office Integrated Data Archive System (MIDAS) Land and Marine Surface Stations Data (Met Office, 2006a; Met Office, 2006b; Met Office, 2006c; Met Office, 2006d; Met Office, 2006e). When recorded hourly, the values were transformed to either mean or total daily values. One additional weather variable was created based on the rainfall data, indicating the

Formatted: Font: English (United Kingdom)

Formatted: Font color: Text 1

Formatted: Font color: Text 1

Formatted: Font color: Text 1

Formatted: Font color: Text 1

Formatted: Font color: Text 1

number of consecutive days without rain. Since weather data was gathered from hundreds of weather stations across the Southwest, one value for each weather variable was calculated as a weighted sum of the recorded values among all weather stations. Each property was assigned to the weather station in the closest proximity and the weight of each weather station was based on the number of properties assigned to it. The more properties a weather station was the closest to (more than any other station), the higher the weight of its recordings (Xenochristou et al., 2020a).

Figure 1 gives a brief overview of the distribution of the six weather variables over the period of the study. Weather in England is characterised by mild temperatures and consistent rainfall all year round. Generally, maximum air temperatures vary between 5°C and 25°C, with very few exceptions, mostly over the winter and summer months (Figure 1). Springs and summers are generally characterised by higher temperatures, increased sunshine hours and lower humidity, although seasonality is not as prominent as in continental climates. Finally, the total amount of rainfall seems to be reduced over the spring and summer months, The presence of rainfall however, which is often found to be the determining factor in water demand forecasting studies, is consistent over all seasons, although it appears to be lower over the winter months.

Previous analysis explored the interactions and correlations between all available explanatory
Previous analysis explored the interactions and correlations between all available explanatory
Previous analysis explored the interactions and correlations between all available explanatory
Previous analysis explored the interactions and correlations between all available explanatory
Previous analysis explored the interactions and correlations between all available explanatory
Previous analysis explored the interactions and correlations between all available explanatory
Previous analysis explored the interactions and correlations between all available explanatory
Previous analysis explored the interactions and correlations between all available explanatory

4 Methodology

4 Methodology

4 Methodology

4 Methodology

4 Methodology

4 Methodology

4 Methodology

4 Methodology

This section describes the main steps of the model development process. These include the selection of the spatial aggregation levels and candidate input variables, as well as the description of the modelling technique and model technical implementation and assessment.

4.1 Spatial Aggregation

Initially, the households are grouped spatially based on their postcodes. This way, it is easy to ensure that properties that are grouped together are actually in close geographical proximity and each property is counted exactly one time. As a result, the following three levels of spatial aggregation are created:

- Network grouping: No grouping criteria are used. Consumption is aggregated among all properties for each day in the data (Network, Figure 2a). Due to errors and inconsistencies, consumption is not available for every property over each day. Therefore

Formatted: Font: 12 pt

Formatted: Font: 12 pt

Formatted: Font: 12 pt

Formatted: Font: 12 pt

Formatted: Font: English (United Kingdom)

Formatted: Font color: Text 1

Formatted: Font color: Text 1

Formatted: Font color: Text 1

Formatted: Font color: Text 1

this group can vary in composition among different days, i.e. include a slightly different collection of properties. The network group consists of 1,056 data points (each data point represents one day), with 64-804 properties in each one, depending on data availability for the corresponding day.

- Area-based grouping: The first part of the postcode (e.g. BA) is used to group the properties into one of six areas. This group consists of 6,336 data points (Areas, Figure 2a), with 1-212 properties in each one (depending on data availability for the corresponding postcode and day). Each data point represents the consumption of an area for one day.
- District-based grouping: The first and second part of the postcode (e.g. BA1) is used to group the properties into districts. This group consists of 76,032 data points (Districts, Figure 2a), with 1-56 properties in each one (depending on data availability for the corresponding postcode and day). Each data point represents the consumption of a district for one day.

The three aggregation levels have a different range in household composition (i.e. the types of households they consist of) among the groups. The smaller (district) groups are a lot more diverse in terms of the types of households they contain, compared to the relatively homogenous network grouping. If there were no gaps in the data and information for all households was available for each day in the dataset, all days would contain information about the same properties. Therefore, no variation would exist when aggregating the whole network. More details regarding the household composition of each aggregation of properties are available in the Supporting Information.

In order to create additional spatial scales, the household group size is set to a fixed number (from 5 to 600), for each postcode and level of spatial aggregation (Figure 2b). Each aggregation level has a set number of household groups for each day (this might slightly vary due to missing data), which is 63 for the district level, 6 for the area level, and 1 for the network level. When the household group size is set to a fixed number, the groups that are smaller than the threshold are excluded from the dataset, whereas the groups that are larger are reduced to the fixed number of properties. When this threshold is increased, the number of data points decreases, as groups with less than the required number of households are removed from the data. The result is nine different spatial scales, comprising of different household group sizes (Figure 2b). The group sizes are set to 5, 10 and 20 for the district groups, 40, 80 and 120 for the area groupings and 200, 400, and 600 for the whole network. The dots in Figure 2b illustrate the number and size of household groups that correspond to each spatial scale, for each day in the data.

4.2 Model Inputs

As it was mentioned in the data section, a variety of input variables became available, including past consumption and weather data as well as postcodes and household characteristics. Based on their nature, the variables were divided into four distinct types:

- Past consumption data: Past consumption data are aggregated temporally at the daily level and spatially at multiple scales. A sliding, 7-day window of past consumption is used as input in order to capture the weekly repetition of demand patterns. This means that for every day in the data, the mean daily consumption for each one of the seven days prior to it was used to make predictions.

Formatted: Font: 12 pt

Formatted: Font: 12 pt

Formatted: Font: 12 pt

Formatted: Font: (Default) Times New Roman, 12 pt

Formatted: Font: (Default) Times New Roman, 12 pt

Formatted: Font: (Default) Times New Roman, 12 pt

Formatted: Font: (Default) Times New Roman, 12 pt

Formatted: Font: (Default) Times New Roman, 12 pt

Formatted: Font: (Default) Times New Roman, 12 pt

- Household characteristics: These refer to the occupancy rate, acorn group, garden size, rateable value, council tax band, and metering status. Since each household group is composed of a variety of households with different characteristics, the percentage of households in each category is used as an explanatory variable, rather than the category itself. For example, for the characteristic 'garden size', there are three possible categories, 'large', 'medium' and 'small'. Each category is used as a continuous explanatory variable in the model, with values varying from zero (0% of households) to one (100% of households). In the case of the garden size, a possible composition for a household group is 30% large gardens, 60% medium gardens and 10% small gardens. Thus, the garden size is represented by three explanatory values (0.30, 0.60, and 0.10), one for each category. The same applies to the rest of the household variables.
- Temporal characteristics: These relate to the season and type of day (working day or weekend/holiday). People tend to have different habits over different times of the year as well as the week, thus temporal variables can be helpful in capturing the time variability of demand.
- Weather: Weather information includes four weather variables, air temperature, sunshine hours, relative humidity, and number of consecutive days without rain. These can capture the weather-dependent variability of demand.

The above four variable types are treated as separate entities in the demand forecasting models, as they have very distinct characteristics that relate to their availability, accessibility, reliability, and thus importance for network operators. Some of the variables are always easily accessible, reliable, and ready to use (temporal characteristics). Others can be expensive to acquire, store, and process, or even inaccurate, especially when they are based on forecasts and estimations (weather and past consumption data). Information about household characteristics can be anywhere in between; some are relatively easily accessible (council tax band, metering status, rateable value, and acorn), whereas others need to be collected through questionnaires and inspections (Xenochristou et al., 2020a).

Eight models with different configurations of the above input variables are tested at each level of spatial aggregation (Table 2). Models 1 to 4 include a combination of past consumption data and other characteristics as input whereas models 5 to 8 are built using only temporal, weather, and household characteristics.

4.3 Gradient Boosting Machines

Previous work (Xenochristou and Kapelan, 2020) focused on comparing a selection of machine learning models for water demand forecasting and identifying the one that achieves the best accuracy. In that case, the models were compared at a certain spatial scale, specifically at the postcode area level. This spatial scale was chosen in order to avoid very small groups of properties that would have interfered with the accuracy of the results but also in order to have enough data points to train and test the model. The results obtained showed that the Gradient Boosting Machine (GBM) method combines high prediction accuracy with ease of implementation hence was chosen for this work.

The idea behind GBMs is to combine a set of weak, base learners in order to create one strong learner. In this study, the base learner is decision trees. The way decision trees work is by dividing the dataset at each branch in a way that maximises entropy, i.e. the homogeneity within each of the split groups. At each branch (node) of the tree, a variable as well as a threshold value are chosen for splitting the dataset. The tree will keep dividing until it reaches a limit, typically defined by the user, such as a maximum tree depth or minimum final node size.

The GBM algorithm uses bagging, as well as boosting in order to achieve the best result. Each tree is trained on a subset of the original data, while at each node of the tree, the best variable for splitting is chosen among a random sample of the input variables (bagging). At each step, one regression tree is built on the residual errors of the previous tree with the aim to

t each step, one regression tree is built on the residual errors of the previous tree with the aim to

4.4 Model Implementation and Assessment

In order to build the model, the dataset is randomly shuffled and divided into a training (70% of the data) and a test (30% of the data) set. The training set is used to train and tune the model for the optimum set of hyperparameters, whereas the test dataset does not participate in the model-building phase and is used to carry an unbiased evaluation of the model's prediction accuracy, based on unseen data. Model training is the process of fitting the model on the training data whereas the tuning step refers to the selection of a set of hyperparameters that are chosen before the training begins. These are important as they define how closely or loosely the model fits the training data. In order to enhance the robustness of the hyperparameter selection process, the performance of the hyperparameter values is tested on multiple subsets of the training data using a 5-fold cross validation process (Zhang, 1993). This means that the training set is divided into five parts and at every iteration, four parts are used for training while one is used to assess the model performance.

The GBM is trained and tuned for the optimum set of hyperparameters using the 'h2o' package (LeDell et al., 2019) written for R (R core team, 2013), which serves as an interface for the 'h2o' machine learning platform (Aiello et al., 2019). Predictions are made for different model configurations, groups of properties, and forecast horizons. The model is retrained and retuned for every change in the input variables, forecast horizon, or spatial aggregation. The automated machine learning capability of 'h2o', called 'automl' (h2o.ai, 2019), is used to identify the optimum set of hyperparameters in each case, using a random search (Bergstra and Bengio, 2012). The high number of hyperparameters that require tuning (nine in total) increases significantly the dimensionality of the search space. Thus, any exhaustive grid search manually implemented by the user would be counter-productive, especially since the aim is to train, tune, and compare a large number of models.

Nine hyperparameters are tuned in this study for the GBM algorithm: the total number of trees that construct the final model (ntrees); the size of the subsample of the training dataset used to

Formatted: Font color: Auto

Formatted: Font: (Default) Times New Roman, Font color: Auto

Field Code Changed

train each tree (sample_rate); the maximum tree depth (max_depth); the number of variables that are sampled and tested for splitting at each node, for the overall model as well as for each tree (col_sample_rate and col_sample_rate_per_tree, respectively); the learning rate (learn_rate) of the algorithm, which is used to reduce the contribution of subsequent trees to the final result; the histogram type used to assist with the splitting selection process (histogram_type); and the minimum requirements for splitting at each node (min_split_improvement and min_rows)
. More information regarding the model hyperparameter can be found in the 'h2o' documentation (h2o.ai, 2019).

After the model is properly trained and tuned, it is used on the test dataset to make predictions for daily consumption 1-7 days into the future. The model performance is assessed based on three criteria, the mean absolute percentage error (MAPE), mean square error (MSE), and R^2 correlation coefficient, as each one of these provided slightly different information. The MAPE is intuitive and independent of the scale of the dependent variable, thus it can be used to compare results from different studies and variables of interest (e.g. per capita consumption and per household consumption). The MSE is sensitive to outliers, while the R^2 shows the variance in the dependent variable that can be explained by changes in the independent variable (Xenochristou , 2019).

Formatted: Font: (Default) Times New Roman

Formatted: Font: 12 pt

5 Results

6.1 Demand forecasting accuracy at different spatial scales

Increasing the level of spatial aggregation consequently decreases the randomness and variability of the water demand signal, making it easier to predict. However, it is unclear by how much. In the following, the relationship between household group size and prediction accuracy is investigated in detail.

First, nine models are trained and tuned for the optimum set of hyperparameters, and consequently assessed for their ability to predict demand for different household group sizes, one day into the future. For comparison purposes, each model is trained with the same input, 7 days of past consumption. Table 3 shows the aggregation level, group size, and number of data points that were used to train each model as well as the results acquired from each one based on three assessment criteria, the MAPE, MSE, and R^2, for the training and test dataset. The results of the hyperparameter tuning process are summarised in the Supporting Information.

According to Table 3, the prediction error (MAPE and MSE) decreases (i.e. improves) as the group size increases. The minimum MAPE corresponds to the largest aggregation, at the network level, with a group size of 600 households, which has an error of 3.2% for the test dataset (Group size = 600, Table 3). The largest MAPE on the other hand (MAPE = 17%) relates to the smallest aggregation scale, at the district level, with a group size of 5 households

(Group size = 5, Table 3). The R^2 value also increases with the group size, but only within the same aggregation level.

However, it is still not clear which point represents the best balance between prediction accuracy and household group size, i.e. at which spatial scale a further increase in group size does not offer a significant reduction in prediction errors. This is depicted in Figure 3, which represents the balance between the MAPE and spatial scale, for the test dataset. According to Figure 3, the model error increases exponentially as the household group size decreases. When everything else remains the same (model structure, input variables), increasing the prediction group size from 40 to 120 households reduces the MAPE by 2.6% (Figure 3). However, for group sizes below ~20 households, the MAPE increases significantly for a rather small decrease in group size. For example, the MAPE increases an additional 7%, from 10% to 17%, for a decrese of 15 households per group (from 20 to 5). On the other hand, for group sizes above ~200 households, the MAPE decreases marginally for a high increase in group size (Figure 3).

6.2 Variable importance at different spatial scales

The three aggregation levels contain different household group sizes, with different ranges in their daily consumption and different amounts of data points (Table 4). In order to avoid increased prediction errors associated with very small groups (<20 households), whilst allowing to create distinct enough group sizes to allow for a meaningful comparison, the minimum group size is set to 20, 60, and 100, for the districts, areas, and network, respectively. The smaller the aggregation level, the smaller the mean group size and the larger the number of data points. In addition, as consumption becomes more erratic and variable for smaller household groups, the range in daily consumption also increases (Table 4).

Results are summarised in Figure 4 and Table 5. Figure 4 shows the prediction accuracy, in terms of MAPE, for predictions 1-7 days ahead, over all days in the data (plots a-c, Figure 4), as well as peak days, i.e. 10% of the days with the highest consumption (plots d-f, Figure 4). Each plot represents one aggregation level (network, area, district) and eight model configurations, with each configuration corresponding to a different set of input variables (Table 2). Table 4 shows the MAPE for each model and each aggregation level, for one as well as seven days into the future, for all days and peak days. The hyperparameter values selected for each model are available in the Supporting Information.

The best performing model for the network level is the one that uses all explanatory variables to make predictions (model 1). When past consumption data is included in the model (models 1-4), temporal characteristics reduce the MAPE by 0.5%, for predictions 1 day ahead (model 3), while weather input further reduces errors by 0.4% (model 2) and household characteristics by 0.1% (model 1). For models 5-8 (no past consumption data), weather input reduces the MAPE by 0.4% (model 7), while household characteristics reduce it by 0.1% (model 6). Adding both household and temporal characteristics (model 5) reduces model errors by 0.9% (Table 5.5).

Although the MAPE value and variance increase for peak days, results are overall very similar. The best performing model (MAPE = 4.6%), for one day lead time, is the one that uses all

Formatted: Font: (Default) Times New Roman

Formatted: Font: (Default) Times New Roman

Formatted: Font: (Default) Times New Roman

Formatted: Font: 12 pt

predictors (model 1). However, for predictions seven days into the future, the model with temporal, household, and weather characteristics (model 5) performs better (MAPE = 6.1%) than the model (model 1) that also incorporates past consumption data (MAPE = 6.4%) (Table 5). Temporal characteristics, on top of past consumption, improve the MAPE by 2.5% (model 3), for one day lead time. Weather input further reduces errors by 0.2% (model 2) and household characteristics by 0.6% (model 1). For models 5-8 however (the ones excluding past consumption data), household and weather input reduce errors by 0.4% (model 6) and 0.1% (model 7), for predictions one day ahead. The combined effect of both of the above reduces the MAPE by 1.3%, a reduction much higher than the simple addition of their individual contributions (model 5). In both cases (all days & peak days), the model that includes only temporal and weather variables (model 7) performs better than the model that includes only past consumption data (Model 4) (Table 5).

As the level of spatial aggregation decreases, the range in errors among the models drastically increases. The best performing model for the areas is still the one that includes all variables (model 1), for all days as well as peak days (Figure 4, (b) and (e)). In this case, temporal, weather, and household characteristics, on top of past consumption data, reduce errors by 0.7%, 0.3%, and 0.1%, respectively, for all days, and 3.5%, 0.2%, and 0%, respectively, for peak days. Weather input for the models without past consumption reduces errors by 0.3% (model 7), for one day lead time, whereas household characteristics reduce it by 1.5% (model 6), for all days (Table 5). The combined effect of both household and weather characteristics outperforms again the mere addition of their individual contributions; the model that includes temporal, household, and weather variables (model 5) has a MAPE of 4.2% for predictions one day ahead (an improvement of 2.1%), an error almost as low as the best performing model (model 1) (Table 5). The same is true for peak days; weather (model 6) and household (model 7) input reduce errors by 1.6% each, whereas the combination of the two contributes to an error reduction of 4.1% (Table 5). Finally, for peak days, the model with temporal and weather input (model 7, MAPE = 9.9%) performs better than the model with past consumption data (model 4, MAPE = 10.7%), for one day lead time.

For the district groups, the MAPE range increases further, varying from 6.7% to 12%, for predictions one day ahead, for all days. In this case, past consumption data and household characteristics offer significant improvements, whereas weather is rather irrelevant (Figure 4c). The model that includes all variables as input (model 1) has once again the best performance (MAPE = 6.7%, for one day lead), although temporal, household, and weather input (model 5) can achieve a similar accuracy (MAPE = 6.8%), for all days in the data. For seven days ahead, models 1 and 5 perform equally well for all days in the data (MAPE = 6.8%), whereas model 5 performs slightly worse (MAPE = 10.3%) compared to model 1 (MAPE = 10.0%) for peak days. Past consumption data (model 3) and household characteristics (model 6), on top of temporal characteristics, reduce errors by 4.9%, from 12.0% to 7.1%, for 1 day lead time (Table 5.5). Weather input (models 2 and 7) offers hardly any benefit to the model for predictions across all days. However, it does improve the MAPE by a maximum of 0.6% on peak days (model 2), for predictions seven days ahead. Finally, the model that uses only weather and temporal characteristics (model 7) has almost double the MAPE for all days (MAPE = 12.0%) and triple for peak days (MAPE = 30.2%), compared to the best performing model (model 1).

Formatted: Font: (Default) Times New Roman

It is worth noting the upward trend of all models that include past consumption as an explanatory factor (models 1-4), as predictions move further into the future. Since water consumption is highly auto-correlated from one day to the next one, predictions for one day ahead are more accurate than seven days ahead. However, adding weather and household input does reduce errors for predictions further into the future. On the other hand, for models 5-8 (no past consumption input), the forecast horizon does not have an effect on the model's output (Figure 5). The result of this is that the best model sometimes shifts depending on the forecast horizon, as models that include past consumption often perform best for one day lead time but are outperformed by the ones that have temporal, household, and weather input for increased lead times (e.g. seven days).

7 Discussion

This paper shows that if everything else stays the same, water demand prediction errors improve for larger aggregations of households, reaching constant prediction accuracy for groups larger than ~200 houses. This is likely due to the fact that as the household group size decreases, water demand becomes more variable as well as more random/erratic, and therefore more difficult to predict. This is illustrated by the level of water demand variability, which is clearly associated with the level of spatial aggregation; smaller groups have a much higher daily water consumption range (80-250 litres/capita/day for the district groups) compared to larger ones (115-175 litres/capita/day for the network grouping). As errors reduce for larger group sizes, the R^2 value increases, but only within the same aggregation level. While the variance in the response variable (i.e. the water consumption) decreases as the group size increases, moving to a higher aggregation level (e.g. from districts to areas) also has a negative effect; grouping together houses that are further away from each other potentially creates less homogenous groups and thus reduces the explanatory value of the predictor variables, in this case past consumption.

This demand variability in smaller household groups can be largely explained by different behaviours and habits and thus results can be improved by adding the right explanatory factors as model inputs.

Past consumption data also became more important as the household group size reduced (Figure 4). Household characteristics are embedded in past consumption, in addition to other factors that can define the consumption behaviour of a certain property or group of properties. Therefore, using past consumption data can be particularly valuable for smaller groups, since it can capture the individual behaviour that relates to the variability in their individual characteristics. This is demonstrated by examining the influence of the explanatory variables for the district areas (Figure 4). When past consumption data is available, household characteristics do not further improve the prediction accuracy of the model. However, when past consumption is not used as model input, a combination of household, weather, and temporal characteristics can adequately be used to characterise and thus predict water demand with the same accuracy. For example, adding weather and household variables on top of past consumption reduced the MAPE a maximum of 1.6% for peak days and district areas whereas for the model that did not include past consumption, adding household and weather characteristics achieved a reduction of 19.7%, from 30% to 10.3%.

The effect of weather became noticeable only for larger groups of properties (Network & Areas, Figure 5), while it is rather irrelevant when attempting to predict consumption for smaller household groups (Districts, Figure 4). Previous studies found that the effect of weather on water consumption varies between households, days and times in the year (Xenochristou et al. 2019a). Out of all households in the dataset, only few of them will alter their consumption behaviour based on the weather and therefore using weather input cannot improve predictions at small levels of spatial aggregation. In these cases, the model would 'learn' based on the majority of the data points, for which weather does not actually have an influence on consumption. However, when aggregating all properties for each day in the data, the effect of weather can be seen in each data point (each day) used to train the model, therefore in this case weather is found to have a (slight) impact on consumption. Notably, the combined contribution of household and weather characteristics in the model was in most cases much higher than their individual contributions. This result confirms further what was already concluded from previous studies (Xenochristou et al., 2020a), that the influence of weather on water consumption is dynamic and it strongly depends on the type of property and residents. Therefore, providing additional context in terms of household characteristics on top of weather information can further improve results.

Finally, implementing more dimensions to the problem, such as the temporal aggregation and model choice would provide more insights into their effect on the results. Here, a GBM model and daily scale is used to compare the forecasting accuracy and variables of interest at different spatial scales. The daily scale allowed to incorporate additional input variables in the model, such as the day of the week, and account for the weekly pattern of water consumption. The GBM model was chosen for its accuracy and ease of implementation, based on previous work that compared the forecasting accuracy of several machine learning models under different scenarios (Xenochristou and Kapelan, 2020). Ideally, all models should be tested under all different scenarios, including different spatial scales, in order to determine the best one for each application. In addition, further work is needed in order to develop a grid of spatial and temporal aggregations of consumption that will demonstrate the limitations and opportunities that arise at each scale. However, including each aspect of the water demand forecasting problem as an unknown variable would increase significantly the dimensionality of the problem. As a result, it would also increase disproportionally the computational and time requirements of the analysis, and equally the processing and understanding of the results. In this case, the model type was considered a fixed (rather than variable) value.

8 Summary and Conclusions

This study explored the effect of the spatial scale on water demand forecasting, both in terms of prediction accuracy and influencing factors. In order to achieve this, multiple models with different input variables were trained on real-life UK daily consumption records for different aggregations of consumption. Initially, three different levels of spatial aggregation were created using the properties' postcode. One group included all the households in the network (up to 804 properties/group) while the other two aggregated the properties in the dataset in 6 areas (up to 262 households/group), or 63 districts (up to 56 households/group). At the same time, three household group sizes were fixed and tested for each aggregation level, varrying from 5 (for the districts) to 600 (for the network) properties per group per day. A Gradient Boosting Machine (GBM) was trained using each of the above configurations and a prediction was made for the water consumption of the same groups, for one day into the future, using only past consumption as an explanatory factor. The purpose of this was to compare the modelling accuracy among models for different spatial scales. After this, different types of model input variables (temporal characteristics, weather data, household characteristics, past consumption) were used in order to improve the prediction accuracy at each level of spatial aggregation (Network, Areas, Districts) and identify the most influential input factors.

The results obtained show the following:

1. The level of spatial aggregation has a direct influence on the demand forecasting accuracy. In general, the higher the spatial scale of household aggregation, the more accurate are demand forecasts. For groups of fewer than 20 households, the prediction error measured via MAPE increases exponentially with a decrease in household group size. On the other hand, for group sizes above approximately 200 households, a further increase in group size only marginally reduces the MAPE.
2. Demand forecasting errors can be reduced by using additional explanatory variables, especially in the case of smaller groups, where the error range varried significantly depending on the input factors used. In this study, the most influential input variables that improved the demand forecasting accuracy varied for different levels of spatial aggregation. Past consumption became more important for smaller aggregations of properties, along with household characteristics, whilst weather data contributed to the model's accuracy only for larger household groups.

Although the effect of different levels of spatial aggregation was investigated in detail in this paper, this was done within a fixed set of environmental conditions. All of the above analysis reflected the consumption of houses in the southwest of England. In a different setting, with different prominent household and resident characteristics, as well as climate, these results could be very different. Although the above methodology could be replicated anywhere where the related data is available, it is important to note that the results could possibly vary.

Acknowledgments

Acknowledgments

Acknowledgments

Acknowledgments

Acknowledgm

This study was funded as part of the Water Informatics Science and Engineering Centre for Doctoral Training (WISE CDT) under a grant from the Engineering and Physical Sciences Research Council (EPSRC), grant number EP/L016214/1.

The data for this study was made available by Wessex Water and is protected under a non-

The data for this study was made available by Wessex Water and is protected under a non-disclosure agreement. Interested parties can ask for data access directly from Wessex Water. The weather data used in this study was collected and became available by the Met Office. This data was provided to the author for research purposes only and is available for purchase by the Met Office.

References

Aiello, S., Eckstrand, E., Fu, A., Landry, M., & Aboyoun, P. (2016). *Machine Learning with R and H2O: Seventh Edition.*

Anele, A. O., Hamam, Y., Abu-Mahfouz, A. M., & Todini, E. (2017). Overview, comparative assessment and recommendations of forecasting models for short-term water demand prediction. *Water (Switzerland)*, *9*(11). doi: 10.3390/w9110887.

Bailey, S. (2019) 'Consumption per head in the UK : 2017', (December), pp. 1–9.

Balling, R. C., Gober, P., & Jones, N. (2008). Sensitivity of residential water consumption to variations in climate: An intraurban analysis of Phoenix, Arizona. *Water Resources Research*, *44*(10), pp. 1–11. doi: 10.1029/2007WR006722.

Brentan, B. M., Meirelles, G., Herrera, M., Luvizotto Jr., E., & Izquierdo, J. (2017). Correlation Analysis of Water Demand and Predictive Variables for Short-Term Forecasting Models. *Mathematical Problems in Engineering*, pp. 1–10. doi: 10.1155/2017/6343625.

CACI Limited (2014). The Acorn User Guide. *The Consumer Classification*, p. 108. Available at: https://acorn.caci.co.uk/downloads/Acorn-User-guide.pdf.

Chen, J., & Boccelli, D. L. (2018). Forecasting Hourly Water Demands With Seasonal Autoregressive Models for Real-Time Application. *Water Resources Research*, *54*(2), pp. 879–894. doi: 10.1002/2017WR022007.

H2O.ai (2019), *AutoML: Automatic Machine Learning*. Available from: http://docs.h2o.ai/h2o/latest-stable/h2o-docs/automl.html [Accessed 06/09/2019]

Herrera, M., Garcia-Diaz, J. C., Izquierdo, J., & Perez-Garcia, R. (2010). Predictive models for forecasting hourly urban water demand. *Journal of Hydrology*, *387*(1–2), pp. 141–150. doi: 10.1016/j.jhydrol.2010.04.005.

House-Peters, L. A., & Chang, H. (2011). Urban water demand modeling: Review of concepts, methods, and organizing principles. *Water Resources Research*, *47*(5). doi: 10.1029/2010WR009624.

House-Peters, L., Pratt, B., & Chang, H. (2010). Effects of Urban spatial structure, sociodemographics, and climate on residential water consumption in Hillsboro, Oregon. *Journal of the American Water Resources Association*, *46*(3), pp. 461–472. doi: 10.1111/j.1752-1688.2009.00415.x.

Hutton, C. J., & Kapelan, Z. (2015). A probabilistic methodology for quantifying, diagnosing and reducing model structural and predictive errors in short term water demand forecasting. *Environmental Modelling and Software*. Elsevier Ltd, 66, pp. 87–97. doi: 10.1016/j.envsoft.2014.12.021.

LeDell, E., Gill, N., Aiello, S., Fu, A., Candel, A., Click, C., Kraljevic, T., Nykodym, T., Aboyoun, P., Kurka, M., Malohlava, M., Rehak, L., Eckstrand, E., Hill, B., Vidrio, S., Jadhawani, S., Wang, A., Peck, R., Wong, W., Gorecki, J., Dowle, M., Tang, Y., DiPerna, L., & H2O.ai (2019), Package 'h2o', R package version 3.26.0.2

Lee, S. J., Wentz, E. A., & Gober, P. (2010). Space-time forecasting using soft geostatistics: A case study in forecasting municipal water demand for Phoenix, Arizona. *Stochastic*

Environmental Research and Risk Assessment, *24*(2), pp. 283–295. doi: 10.1007/s00477-009-0317-z.

Maheepala, S., Blackmore, J., Diaper, C., Moglia, M., Sharma, A., & Kenway, S. (2011). *Integrated Urban Water Management Planning Manual*.

Matos, C., Teixeira, C. A., Bento, R., Varajao, R., & Bentes, I. (2014). An exploratory study on the influence of socio-demographic characteristics on water end uses inside buildings. *Science of the Total Environment*. Elsevier B.V., 466–467, pp. 467–474. doi: 10.1016/j.scitotenv.2013.07.036.

Met Office, N. B. A. C., 2006e. *MIDAS: U.K. Soil Temperature Data,* s.l.: s.n.

Met Office, N. B. A. D. C., 2006a. *MIDAS: U.K. Daily Rainfall Data,* s.l.: s.n.

Met Office, N. B. A. D. C., 2006b. MIDAS: U.K. Daily Temperature Data, s.l.: s.n.

Met Office, N. B. A. D. C., 2006c. MIDAS: U.K. Daily Weather Observation Data, s.l.: s.n.

Met Office, N. B. A. D. C., 2006d. MIDAS: U.K. Hourly Weather Observation Data, s.l.: s.n.

Ofwat (2017). *PR19. Appendix 2: Delivering outcomes for customers.*

Ofwat (2018). *Reporting Guidance Leakage,* pp. 1–44. Available at: https://www.ofwat.gov.uk/publication/reporting-guidance-leakage/.

Ofwat (2019). *Ofwat's forward programme 2019-20.*

Parker, J. M. & Wilby, R. L. (2013). Quantifying Household Water Demand: A Review of Theory and Practice in the UK. *Water Resources Management*, 27(4), pp. 981–1011. doi: 10.1007/s11269-012-0190-2.

Polebitski, A. S., & Palmer, R. N. (2010). Seasonal Residential Water Demand Forecasting for Census Tracts. *Journal of Water Resources Planning and Management*, *136*(1), pp. 27–36.

R Core Team (2013). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

Rathnayaka, K., Malano, H., Arora, M., George, B., Maheepala, S., & Nawarathna, B. (2017a). Prediction of urban residential end-use water demands by integrating known and unknown water demand drivers at multiple scales I: Model development. *Resources, Conservation and Recycling*. Elsevier B.V., 117, pp. 85–92. doi: 10.1016/j.resconrec.2016.11.014.

Rathnayaka, K., Malano, H., Arora, M., George, B., Maheepala, S., & Nawarathna, B. (2017b). Prediction of urban residential end-use water demands by integrating known and unknown water demand drivers at multiple scales II: Model application and validation. *Resources, Conservation and Recycling*. Elsevier B.V., 117, pp. 85–92. doi: 10.1016/j.resconrec.2016.11.014.

Romano, M. and Kapelan, Z. (2014) 'Adaptive water demand forecasting for near real-time management of smart water distribution systems', *Environmental Modelling and Software*. Elsevier Ltd, 60, pp. 265–276. doi: 10.1016/j.envsoft.2014.06.016.

Royal Mail (2012). *Postcodes and Addresses explained*, available from: https://web.archive.org/web/20120716211307/http://www.royalmail.com/delivery/mail-advice/postcodes-addresses-explained [Accessed 05/09/2019].

Ulanicki, B., AbdelMeguid, H., Bounds, P., & Patel, R. (2009). Pressure control in district metering areas with boundary and internal pressure reducing valves. *Geotechnical Special Publication*, (187), pp. 691–703. doi: 10.1061/41023(337)58.

Xenochristou, M., Hutton, C., Hofman, J. & Kapelan, Z. (2019a). Using smart demand-metering data and customer characteristics to investigate the influence of weather on water consumption in the UK. *Journal of Water Resources Planning and Management*, doi: 10.1061/(ASCE)WR.1943-5452.0001148.

Xenochristou, M., Hutton, C., Hofman, J., & Kapelan, Z. (2019b). A new approach to forecasting household water consumption (submitted).

Xenochristou, M., & Kapelan, Z. (2019). An ensemble stacked model with bias correction for improved water demand forecasting (submitted).

Zhang, P. (1993). Model Selection Via Multifold Cross Validation. *The Annals of Statistics*, 21(1), pp. 299–313. URL : http://www.jstor.org/stable/303559

Zubaidi, S. L., Gharghan, S. K., Dooley, J., Alkhaddar, R. M., Abdellatif, M. (2018). Short-Term Urban Water Demand Prediction Considering Weather Factors. *Water Resources Management*, *32*(14), pp. 4527–4542. doi: 10.1007/s11269-018-2061-y.