Smart-Meter Enabled Estimation and Prediction of Outdoor Residential Water Consumption

by

Valerie Platsko

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Master of Mathematics
in
Computer Science

Waterloo, Ontario, Canada, 2018

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Abstract

Smart meter technology allows frequent measurements of water consumption at a household level. This greater availability of data allows improved analysis of patterns of residential water consumption, which is important for demand management and targeting conservation efforts. The dataset in this thesis includes 8,000 single family residences in Abbotsford, British Columbia from 2012–2013, and contains hourly measurements of water consumption recorded by smart meters installed in 2010. This work focuses on identifying outdoor consumption due to its contribution to peak demand during the summer, which is important because of concerns about strain on infrastructure in Abbotsford. This research shows that outdoor water consumption can be robustly identified from hourly measurement of total water consumption by determining an upper threshold on plausible indoor usage, and that this estimated outdoor water consumption is consistent with seasonal patterns of water consumption identified in previous work, with the timing of restrictions on outdoor watering, and with household size. The research also includes regression tree-based models for predicting next-hour water consumption, however the predictability of this consumption is limited. In contrast to previous work, there is little correlation between outdoor consumption and demographic factors such as income. Outdoor consumption shows a large amount of individual variability, with 8.6% of households accounting for 50% of the total outdoor usage. This limits the predictability of outdoor consumption, but also highlights the importance of identifying this consumption for each household to allow for targeted conservation efforts.

Acknowledgements

First, I'm grateful to my advisor, Professor Peter van Beek, for his advice, patience, and encouragement. Thanks also to my labmates Steven Wang and Irish Medina. In particular, I'd like to thank Steven for his considerable work preprocessing and cleaning the water consumption and weather data, and for providing information about the datasets. Finally, I'd like to thank Professors Kate Larson and Jesse Hoey, for reading and providing helpful feedback on this thesis.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

In this chapter, I informally introduce the problems in my thesis related to identifying and predicting outdoor residential water consumption. I also summarize the contributions of the thesis and the organization of the thesis.

1.1 Residential Water Demand and Outdoor Water Consumption

Accurate knowledge about water demand is important for a variety of reasons including short-term operating decisions for water utilities, long-term planning on the part of municipalities, decisions about when to enact watering restrictions, and for targeting conservation messages. In particular, it is beneficial to have information both about the amount of water consumed and about how that water is consumed [1]. The amount of water consumed and the timing of that consumption determines the strain on infrastructure, while the end-uses of that consumption can be used to target conservation efforts and to make estimates about what amount of reduction might be realistic [2].

This thesis focuses on residential water consumption in Abbotsford, British Columbia. The main dataset is one year of water consumption measurements for each household, recorded from September 2012 to August 2013. The dataset includes hourly water consumption measurements for each customer, recorded by the smart meters installed by the municipality in 2010. The increased frequency of water measurements allowed by smart water meters (see Chapter 2) allows more detailed methods of analysis than water meter readings collected at a monthly or bimonthly frequency for billing. Abbotsford is most concerned about high levels of water consumption during the summer, because increased demand occurring at the same time contributes to strain on infrastructure. This thesis focuses on identifying, predicting, and explaining outdoor water consumption (such as water used for irrigation) during the summer months. Understanding this consumption is important because of its contribution to periods of high total water demand.

1.2 Organization and Overview

In Chapter 2, I discuss general background related to water consumption and to the machine learning models used for prediction. I also outline the problems covered in this thesis in more detail. In Chapter 3, I discuss related work in identifying individual uses of water from a single measurement of total consumption, forecasting water consumption, and the determinants of both total and outdoor consumption. Chapter 4 discusses the datasets used in the rest of the work.

The next three chapters discuss the main contributions of my thesis:

- Chapter 5 discusses an approach to estimating hourly water consumption, adapted from previous work [2]. This approach takes advantage of the fact that outdoor uses of water such as irrigation are high-volume compared to indoor use, and involves identifying an upper threshold indoor water consumption, such that an hourly consumption measurement past the threshold indicates probable outdoor consumption. In contrast to the previous work, potential alternative thresholds are evaluated using qualitative and quantitative evidence to determine the ideal threshold.

- Chapter 6 discusses models for predicting next-hour outdoor water consumption (estimated using the method above). The models are based on ensembles of regression trees to allow some interpretability, and are the first models for predicting outdoor water consumption at a fine temporal scale (hourly) and a fine spatial scale (small neighbourhoods).

- Chapter 7 includes an analysis of the factors affecting outdoor water consumption over the whole summer. In contrast to previous work, demographic factors such as income are not strongly correlated with outdoor water consumption. I also find significant individual variability in outdoor water consumption, with a small number of households consuming most of the water used for outdoor purposes. Previous work [3, 4] that focused on analysing outdoor water consumption required specially installed high-resolution water meters that limit the viable sample sizes. The approach to separating outdoor water consumption described previously allows this analysis using only existing water meters, and therefore it can be performed on a larger number of households.

Finally, Chapter 8 contains some concluding remarks and discussion of future work.

Chapter 2

Background

In this chapter, I briefly review the necessary background in residential water consumption generally, the concerns about water consumption in Abbotsford, smart water meters, time series forecasting as supervised learning, and ensembles of regression trees.

2.1 Residential Water Consumption and Forecasting

This thesis focuses on explaining and predicting single-family outdoor residential water consumption. This section covers relevant facts about residential water consumption, and about how forecasts are developed and for what purposes they are used. For technical details about various types of forecasting, see Chapter 3.

Managing residential water demand is important because of strain on infrastructure, growing populations, and concerns about water pressure [1]. Additionally, residential water is a significant part of the treated water supplied by municipal water utilities [5]. In particular, it is important to manage *peak demand*, which is the largest amount of water required in a fixed period of time (such as a day) [6] in situations where the concern is not the total amount of water used, but potential strain on the water-delivery infrastructure's ability to deliver the required amount of water at one time.

Cole and Stewart [2] note that improved water demand forecasting will require accurate knowledge of the timing, location, and purposes of water use. *When* water is used depends primarily on three cycles of different frequencies [7]. Residential water consumption has a daily pattern in which consumption peaks in the morning and in the evening, which can be attributed to activities such as bathing and cooking, before and after work. There is relatively little consumption overnight. There is a also weekly cycle, with different patterns of consumption on weekends and weekdays. Additionally, in many climates there is a seasonal pattern, with larger amounts of water consumption during the summer months [8]. These patterns are particularly important for managing peak demand because the timing of water consumption, rather than only the total amount, is relevant in that case.

How and where water is consumed are also relevant for demand management and making predictions or inferences about future water consumption. Outdoor uses of water such as irrigation have typically been considered discretionary [4], and can be more easily reduced in response to water shortages than necessary indoor end uses such as toilet flushing. In comparison, indoor consumption is largely used for basic needs and primarily determined by household size [8], although there are indoor end-uses such as showering which are also behaviourally-influenced [4]. Where there are concerns about water demand, it is also important to know the purposes (end-uses) of that consumption, in order to target conservation efforts and estimate the amount that water consumption could be reasonably reduced [2].

In addition to understanding water consumption, it is often useful to be able to estimate the total amount of future consumption. Forecasts may be developed for various purposes such as day-to-day management of water systems or for longer term planning and the ideal forecasting method depends on the purpose of the forecast, its required accuracy, and the resources (including data) which are available [6]. In practice, very simple forecasting methods such as multiplying per-capita consumption by projected population growth are common [9]. Forecasts may be of different periodicities (for example, daily, monthly, or annual predictions) and have different time horizons (for example, a monthly forecast of the next 12 months). Donkor et al.'s [9] survey on forecasting methods describes three basic categories of forecasts depending on the time horizon: operational forecasts for management of water systems (which typically have hourly, daily, monthly, or annual periodicity), tactical forecasts for revenue estimation and investment planning (monthly or annual periodicity), and strategic forecasts for planning major infrastructure expansion (annual periodicity and multi-year time horizon). The accuracy of the forecasts required for any of these purposes depends on many factors such as the likely strain on infrastructure, the amount of water available, and the costs of infrastructure expansion [6].

See [8] for an overview of patterns in residential water use in North America, and [6] for an overview of the forecasting methods typically used by water utilities and the trade-offs involved. Also see [10] for an overview of how urban water consumption is modelled.

2.2 Summer Water Demand in Abbotsford

The water consumption data in this thesis is from Abbotsford, British Columbia. During the data collection period (September 2012 to August 2013) the city had ongoing conservation measures to reduce peak day water demand. The peak day water demand in Abbotsford typically occurs in July or August due to increased outdoor water consumption, and conservation measures include rebate programs, sprinkling restrictions, seasonal water rates, and education about efficient irrigation [11]. The motivation for reducing peak day consumption is to avoid requiring infrastructure expansion which has both financial and environmental costs [12].

During July and August 2013, the city of Abbotsford enacted watering restrictions, as they have in most years since 1995, in order to reduce the peak day demand [12]. Lawn irrigation was permitted on two days per week per household from 6:00 am to 8:00 am, determined by the street address. Even-numbered houses were permitted to water their lawns on Wednesdays and Saturdays, and odd-numbered houses were permitted to water their lawns on Thursdays and Sundays. These restrictions only applied to using sprinklers for automatic irrigation. Manual watering of gardens and trees with a spring-loaded hose and filling pools were permitted at any time.

2.3 Smart Meters

Smart water meters were installed in Abbotsford beginning in 2010, with the goals of detecting leaks, reducing meter reading costs, and collecting data for targeting conservation initiatives [13]. Smart meters are water meters that allowing the logging, storage, and transmission of water consumption measurements [14]. Typically smart meters record household water consumption at a higher frequency than the monthly or bi-monthly frequencies required for billing, and measurements every 15 minutes or every hour are common [14]. This increased temporal resolution allows better demand management by showing when water is consumed, which is important for managing peak demand [15]. Smart meters are useful for both providing real-time feedback about water consumption habits to consumers in order to encourage conservation, as well as useful to water utilities for water consumption purposes [16].

The water consumption measurements recorded by smart water meters can be processed and analysed in the same way as any other signal [17]. This is useful for short-term demand forecasting as well as giving better insight into the end-uses of water than allowed by measurements of water consumption at the frequency required for billing [15]. (See Chapter 3 for methods of identifying the end-uses of water from household-level measurements of total water consumption.)

See [14] for a basic description of how smart meters function and [15] and [16] for a summary of the ways smart meters are used to encourage water conservation and for management of water systems.

2.4 Times Series Forecasting as Supervised Learning

The water consumption data collected from Smart Meters can be thought of as a time series: a series of measurements of a quantity at regular points in time. The measurement of the variable y at time t is denoted y_t:

$$y_1, y_2, ...y_{t-2}, y_{t-1}, y_t.$$

Often, the goal is to predict future values of the time series. Predicted values of the time series are written as $\hat{y}_t$. Predictions for a time series may also take into account other variables, $x_1, ..., x_n$. In the time series literature, previous values of the series y_{t-k} are referred to as *lagged values* and $x_1, ..., x_n$ are referred to as *exogenous variables*.

Time series prediction can be formulated as a supervised learning problem [18]. Supervised learning takes a set of *features* and fits a model which maps the features to *outputs*. The rest of this thesis uses the terminology common in the machine learning literature, referring to both lagged values of y and exogenous variables as features. Given the past k values of y, the relationship of the output to the features can be written as:

$$y_t = f(y_{t-1}...y_{t-n}, x_1, ..., x_n).$$

Where f is a function that maps the features to the output y_t. The formalization above is given for simplicity, but the model may use non-consecutive previous values of y. Once the task is modelled as a supervised learning problem, y_t can be predicted using standard machine learning techniques. The goal of supervised learning is to find a function $\hat{f}$ which approximates f such that the differences between the true values y_t and the predicted values $\hat{y}_t$ are minimized:

$$\hat{y}_t = \hat{f}(y_{t-1}...y_{t-n}, x_1, ..., x_n).$$

The vector of all of the features, $y_{t-1}...y_{t-n}, x_1, ..., x_n$, is referred to as an *observation* and is subsequently denoted as X for simplicity:

$$\hat{y} = \hat{f}(X).$$

It is possible to choose $\hat{f}$ such that it closely predicts correct outputs for the data used to fit the model, but does not generalize well to unseen data. To avoid *overfitting* to the data used to fit the model, the observations are typically divided into at least two sets. The *training* set is used to fit the model, and the *testing* set is used to evaluate its performance. When a limited number of observations are available, *cross-validation* is typically used to evaluate the model, which involves partitioning the dataset into several *folds*, and using one fold as test data and the rest as training data on each iteration.

See [18] for more background on supervised learning for time series prediction, [19] for background on supervised learning generally, and [20] for background on evaluating machine learning algorithms.

2.5 Regression Trees

Regression trees are an interpretable machine learning method that involves partitioning the feature space into rectangular regions and predicting a simple function for each region. The description given here is of regression trees built using the CART algorithm [21], and the explanation is adapted from [19], which can be referred to for more detail.

When building a regression tree, the feature space is divided into regions by creating recursive binary partitions over a single variable at a time. See Figure 2.1 for an example of this partitioning. The value in each region R_m is predicted as a constant c_m. Consider the case with two continuous-valued features, x_1 and x_2 as a simple example of this partitioning. Figure 2.1 shows a potential partition of the feature space and its associated tree. Each time the input is partitioned, it is split into two halves, with one half containing the observations $\{X|x_j < s_i\}$ and the other containing $\{X|x_j \geq s_i\}$.

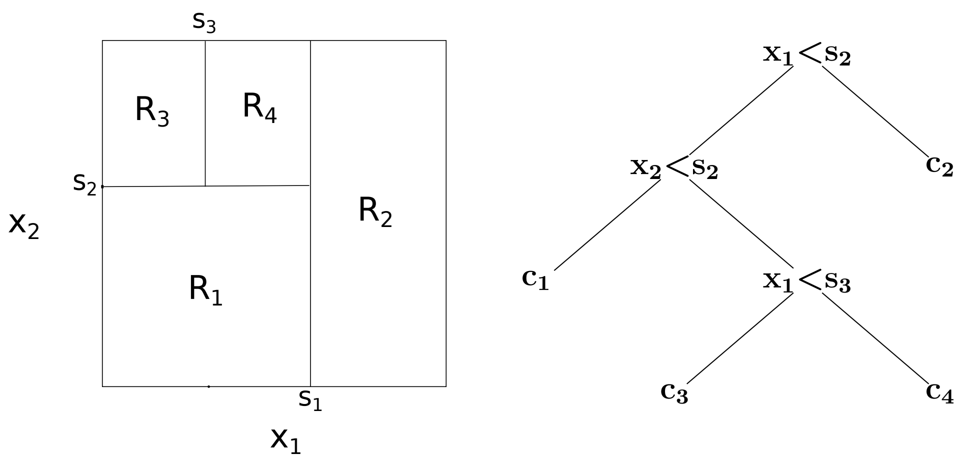

Figure 2.1: Example of input space partitioning for regression trees

The predicted value for an observation X is simply a constant associated with the region that contains it:

$$\hat{f}(X) = \sum_{m=1}^{M} c_m I(X \in R_m).$$

Where $I(X \in R_m)$ has value 1 when X is contained in R_m and value 0 otherwise. Using the sum of squares error for the predictions in a single region, $\sum_{i\in\{i|X_i\in R_m\}}(y_i - c_m)^2$, the optimal value for c_m is the mean value of the observations that fall in the region,

$$c_m = \text{mean}(y_i \mid X_i \in R_m).$$

The regions are partitioned using a greedy algorithm, splitting based on a single feature at a time. Each time a region is partitioned, the variable x_j and the threshold value s at which to split are chosen to minimize the sum of errors in the new regions,

$$\min_{j,s}\left(\sum_{i\in\{i|x_{ij}<s\}}(y_i - c_{j,s})^2 + \sum_{i\in\{i|x_{ij}\geq s\}}(y_i - c_{j,s})^2\right).$$

Regions are partitioned recursively until some stopping condition is met. Stopping criteria can include a maximum depth for the tree or a minimum number of observations per region.

While regression trees are visually interpretable, it is also possible to quantify the influence that each feature has on reducing the error of the prediction. For each internal node of the tree, t, the reduction in the error by splitting the initial region R_1 into new regions R_2 and R_3 (compared to the error if R_1 had not been partitioned) can be calculated as:

$$i_j^2 = \sum_{i\in\{i|X_i\in R_1\}}(y_i - c_1)^2 - \sum_{i\in\{i|X_i\in R_2\}}(y_i - c_2)^2 - \sum_{i\in\{i|X_i\in R_3\}}(y_i - c_3)^2.$$

This value is averaged over all internal nodes $t \in J$ in a tree T that split on a feature x_j to give the *variable importance* for x_j,

$$I_j^2(T) = \sum_{t\in J} i_t^2.$$

2.6 Ensembles of Regression Trees

A single regression tree is easy to interpret visually, but using multiple trees to make a prediction can improve the prediction accuracy. In this section I describe boosted regression trees using the LSBoost algorithm [22], which involves fitting subsequent trees to the errors of the trees already in the model in order to iteratively improve performance. As in the previous section, the description is adapted from [19].

In the LSBoost algorithm applied to regression trees, at each step, a regression tree is fit to the *residuals*, $r_i = y_i - \hat{y}_i$ of the model at the previous step. The basic algorithm for fitting an ensemble model with B regression trees is as follows:

1. Set $\hat{f}(x) = 0$ and $r_i = y_i$ for all i
2. For b=1 to B:

    Fit a tree $\hat{f}^b$ to the observations, using r_i as the output variable
    Update $\hat{f}(X) = \hat{f}(X) + \hat{f}^b(X)$

Update the residuals $r_i = r_i - \hat{f}^b(X_i)$

3. The final model is $\hat{f}(X) = \sum_{b=1}^{B} \hat{f}^b(X)$

While ensembles of regression trees are less easily visually interpreted than individual regression trees, the variable importances for the ensemble can be calculated by averaging the variable importances for each tree T_b in the ensemble:

$$I_j^2 = \frac{1}{B} \sum_{b=1}^{B} I_j^2(T_b)$$

2.7 Problem Description

As discussed previously, outdoor water consumption is a major component of peak usage. Additionally, reducing outdoor consumption is an important part of reducing peak demand because water consumption for irrigation is more discretionary than most indoor end-uses of water. Because Abbotsford is concerned with strain on infrastructure, and because water demand is highest in July and August, I focus primarily on analysing outdoor water consumption in these months. The main dataset in this thesis is hourly water consumption data for 2012 and 2013 (see Chapter 4) which was recorded by the smart water meters installed in Abbotsford beginning in 2010. There are three problems considered: identifying outdoor water consumption, predicting next-hour outdoor water consumption, and explaining the determinants of this consumption over the entire summer.

The first problem is identifying outdoor water consumption. Although various methods of separating the end-uses of water exist (see Chapter 3) the hourly resolution of the water consumption measurements in the dataset prevent the use of more complicated methods of disaggregation. I adapt Cole and Stewart's [2] work on identifying outdoor water consumption from hourly measurements. This method involves finding a threshold for hourly usage past which it is implausible that all of the usage is for indoor purposes. I validate this threshold based primarily on the cyclical patterns of water consumption discussed in this chapter, and on water consumption patterns in relation to Abbotsford's watering restrictions.

After the outdoor consumption is identified, it can be treated as time series data and used for developing a forecast of future outdoor water consumption. I develop models to predict the next-hour outdoor water consumption for small neighbourhoods, based on previous water consumption and on demographic data. Hourly forecasts are primarily useful for operational management of water systems. The models are based on ensembles of regression trees, in order to allow them to be interpreted.

Lastly, although the predictive models can allow some insight into the determinants of water consumption at the hourly level, these determinants may vary at different time scales. I also analyse the relationships between outdoor water consumption over the whole summer and demographic and household variables (such as income, lot size, and the presence of a swimming pool), as well as discuss the variability of water consumption between consumers.

2.8 Summary

In this chapter, I described the necessary background related to residential water consumption and supervised learning. I also described the problems covered in the rest of this thesis in more detail.

In the next chapter, I present related work on the three problems described in this chapter: identifying outdoor water consumption, forecasting future water consumption, and explaining this outdoor water consumption.

Chapter 3

Related Work

In this chapter I discuss the previous work related to the problems in this thesis: identifying, predicting, and explaining outdoor water consumption. Because not all of the prediction and explanatory work is focused exclusively on outdoor prediction, I also discuss prediction and explanation of total water consumption where relevant.

3.1 Disaggregating Household Water Consumption

Non-intrusive methods of recording water consumption, such as smart water meters or traditional metering, produce a single measurement per time period per meter. This has the drawback that, while total consumption is known, it is not known which uses contribute to that total or when those uses occur. The obvious solution for this is to install separate meters for each water-consuming appliance, but this is costly, time-consuming, and less acceptable to consumers [1]. A second approach, typically called non-intrusive disaggregation, is to take a single (or a smaller number) of meters, typically outside the house, and from the patterns of consumption, separate different end-uses or fixture categories from each other algorithmically. The approaches possible vary by the recording frequency of the meter used. High frequency recordings (such as multiple measurements per minute) allow much more complex approaches than is possible with data typically collected by smart meters in practice. Fifteen minute or hourly frequency of measurements is more common, in order to reduce storage and transmission costs [14].

Various methods of disaggregation have been developed, depending on the type of meter used for recording. The most common type of meters used to collect data for disaggregation are flow meters [23], which measure the volume of water consumption. Typically flow meters require measurements at 5 second intervals and categorize end-uses by classifying distinct flow patterns from different end-use events [1]. There are various tools available for this kind of disaggregation such as TraceWizard [24], and Identiflow [25]. While flow meters are the most common because they may be installed outside of the house, other methods of sensing exist. These include acoustic [26] and pressure sensing [27, 28], but these both require higher-frequency sampling than do flow meters, and therefore have higher storage and processing cost. In addition, combined approaches exist [29, 30].

3.1.1 Identifying Seasonal Water Consumption

All of the above systems require relatively high-resolution data, and, other than flow meters, additional equipment beyond what is already installed by water utilities for billing purposes. However, for less detailed disaggregation, simpler approaches may be sufficient. In particular, it is possible to estimate outdoor water consumption more easily because it is high-volume in comparison to indoor uses [1]. Although this two-category disaggregation is less detailed than an appliance-level disaggregation, it can still give important insights into how water is being used and how it can be conserved. Outdoor consumption is a good target for conservation efforts because, unlike many categories of indoor use where consumption depends on appliance efficiency, such as clothes washing and toilet flushing, the amount of water used outdoors can be more easily reduced by consumer choices.

Many previous methods of identifying outdoor water consumption have used the monthly or bi-monthly data typically used for billing. They involve subtracting *base use*, defined as monthly usage during the winter, from consumption measurements during the summer in order to estimate outdoor consumption [31–34]. This requires the assumption that indoor use has no seasonal pattern, which is not true for all end uses, such as shower water consumption [35]. Gato et al. [36] introduce a variation on this approach, which uses temperature and rainfall thresholds to identify base use as weather-insensitive use.

Related methods involve using linear regression models to find the sensitivity of consumption to temperature variation, with the assumption that temperature-sensitive use is likely to be primarily outdoor consumption [37, 38]. This gives an estimate of how outdoor consumption is distributed spatially, but does not provide a real disaggregation of consumption.

Castledine et al. [39] produce an estimated disaggregation of outdoor water consumption based on clustering daily water usage and taking into account the presence of watering restrictions. Cole and Stewart [2] estimate outdoor consumption by establishing an hourly threshold past which consumption could not reasonably be primarily for indoor uses. This is the only previous method of disaggregation I am aware of which produces an hourly estimate at the household level. The approach I take is similar, although I establish and test the threshold differently (see Chapter 5).

3.2 Forecasting Water Consumption

Forecasting water consumption is important for various purposes. Short-term forecasts are used primarily for operation and management, while longer-term forecasts are used for planning and infrastructure design [40]. House-Peters et al. [10] suggest that it is important to develop reliable demand forecast models, especially for peak demand. Some previous work focuses on predicting peak consumption at the weekly level [40,41]. However, despite outdoor consumption being an important determinant of the variability in peak demand, few studies have predicted it directly. Taylor et al. [42] predict outdoor water consumption at a monthly timescale for various cities in Australia, but no other previous work that I am aware of predicts outdoor water consumption at a finer temporal or spatial scale.

I discuss relatively short-term (monthly and sub-monthly) prediction methods for *total* water consumption below, with an emphasis on the size of the area for which forecasting is being performed, the temporal scale of the forecast, the modelling method, and the features used. Table 3.1 shows a non-exhaustive comparison for previous work along these axes. See [9] for a more complete review of forecasting approaches.

The majority of previous work produces forecasts for an entire city. It is therefore not directly comparable to this work because larger areas tend to have more predictable water consumption [43]. However, some previous work forecasts at somewhat smaller spatial scales. Herrera et al. [44] predict consumption for a hydraulic sector containing approximately 5000 consumers. Adamowski and Karapataki [41] predict over neighbourhoods of unspecified size, and Jain et al. [45] predict consumption for a university campus with approximately 6,500 students. The most similar work in scale to this work is by Walker et al. [46], which predicts hourly water consumption at the individual household level.

Initial work on forecasting water consumption used time series or regression models. More recent work has used a variety of machine learning models, including artificial neural networks (ANN), support vector regression (SVR), and random forests. Herrera et al. [44] find that random forest models produce comparable results to other methods such as SVR and multivariate adaptive regression splines (MARS), and superior results to the ANN models evaluated.

In summary, there is no work on explicitly predicting outdoor consumption which uses a fine temporal scale (such as hourly) or produces forecasts for a small area. While previous work on predicting total water consumption includes forecast models at a wide variety of temporal scales, the majority of this work uses large spatial scales (such as an entire city or large neighbourhood). The previous work on predicting short-term water consumption at a smaller spatial scale (household level) has limited accuracy. The models I discuss in Chapter 6 predict outdoor consumption at both a much finer temporal scale (hourly) and a finer spatial scale (small neighbourhoods) than previous work on predicting outdoor consumption.

Table 3.1: Previous work on forecasting total water consumption

Paper	Spatial Scale	Temporal Scale	Model Type	Model Variables
Nasseri et al. [47]	city level	monthly	genetic programming and extended Kahlman filters	previous consumption values
Walker et al. [46]	household	hourly	ANN	previous consumption values, averages of previous consumption values, hour of day
Cutore et al. [48]	large neighbourhood (population 50,000)	daily	SCEM-UA ANN	previous consumption values, day of the week
Herrera et al. [44]	hydraulic sector	hourly	ANN, projection pursuit regression, multivariate adaptive regression splines, SVR, random forests, weighted pattern-based model	previous consumption values, day of week, climatic variables
Jain et al. [45]	university campus	weekly	ANN, linear regression, time series models	previous consumption values, rainfall, air temperature
Zhou et al. [49]	water supply zone	daily and hourly	time series models, disaggregation	previous consumption values, rainfall, air temperature
Bougadis et al. [40]	suburb	weekly peak	linear regression, multiple regression, time series, ANN	previous consumption values, rainfall, air temperature
Caiado [50]	country	daily (multistep)	time series models	previous consumption values
Maidment and Parzen [51]	city	monthly	cascaded time series models	previous consumption values, climate, population
Praskievicz and Chang [52]	city level	monthly	time series	previous consumption values, climate
Wong et al. [53]	city level	daily	statistical model	previous consumption values, climate, calendar variables
Miaou [54]	city level	monthly	nonlinear regression models	previous consumption values, climatic variables
Adamowski and Karapataki [41]	neighbourhood	weekly (peak)	regression models, ANNs	previous consumption values, climatic variables

3.3 Explaining Outdoor Residential Water Consumption

Identifying the determinants of outdoor water consumption is important for demand management purposes [2]. In this section I describe the previous work on explaining the determinants of outdoor water consumption. Some previous work has focused explicitly on outdoor consumption, and other work has used the sensitivity to temperature as an indicator of likely outdoor consumption, and then determined the factors that contribute to this sensitivity. I also include some work on the determinants of total water consumption where relevant.

The influence of various factors may depend on the timescale studied. Miaou [54] finds that weather is the primarily determinant of short-term variation in water consumption, while demographic characteristics are important at longer timescales.

3.3.1 Weather

Previous work has shown that water consumption is dependent on weather. Although generally this work analyses total consumption, outdoor use is thought to be the primary driver of this consumption. Results on the influence of weather have varied by location. For example, Mayer et al. [8] found greater per-capita water consumption in locations with hotter climates.

In addition, the influence of weather is not likely to be constant across the year. For example, Akuoku et al. [55] find that total water consumption is only dependant on temperature past a threshold value. Similar findings were reported by Gato et al. [36] and Maidment and Miaou [43].

The impact of weather on water consumption also appears to vary by location and the particular water usage patterns in the city studies. Balling et al. [56] find that residential water consumption in Pheonix is not as variable as expected in response to weather conditions, which they hypothesize is because of appliances such as automatic sprinklers which are not adjusted in response to changing weather conditions.

3.3.2 Income

Previous work has generally shown outdoor water consumption to be positively correlated with higher income (although this was not true in the Abbotsford dataset; see Chapter 7). Loh et al.'s [3] end-use study in Australia finds that income is highly correlated with outdoor consumption but not with indoor consumption. Additionally, Syme et al. [34] similarly find that outdoor consumption is correlated with income. However, Willis et al. [4] show that within middle-to-upper income houses, no significant differences in outdoor water consumption are detected, so this may vary based on the range of incomes studied. This previous work on explaining outdoor consumption used specially installed higher-frequency smart meters [3, 4], which limits the sample sizes that can be used, or produced a rough estimate of summer outdoor water demand from billing data by subtracting winter consumption [34]. Additionally, Corbella et al. [57] suggest that the dependence on income may not be direct sensitivity to price, but rather vary based on lifestyle factors such as the ownership of water-using appliances and pools. Similarly, Harlan et al. [58] find that the influence of income on total water consumption was not significant once other factors such as irrigable outdoor space and house size were considered. Mayer et al. [8] show that for various North American cities, the end-use of water most correlated with income is irrigation.

Some additional work has studied the demographic factors that related to the sensitivity of total water consumption to changes in temperature and precipitation. This sensitivity to temperature is used as a measure of outdoor consumption. Balling et al. [37] find that in Phoenix census tracts with a large percentage of high-income households are most sensitive to changes in weather conditions. Polebitski et al. [59] also show that for their study in Seattle, the ratio of summer to winter consumption, as well as total summer consumption are significantly influenced by income.

3.3.3 Property and Building Factors

In addition to demographic factors, previous work has shown that water consumption is related to property factors such as lot size, the amount of irrigable outdoor space, and the type of vegetation. In contrast to most of the previous work, there is not a strong relationship between lot size and outdoor water consumption in this research (see Chapter 3).

Breyer et al. [38] compare the temperature sensitivity of water consumption in Portland to that in Phoenix and find that water consumption is most sensitive to outdoor space in Portland, and most sensitive to vegetation type in Phoenix. Willis et al. [4] find that irrigation increases with lot size. In contrast, Loh and Coghlan's end use study in Western Australia found no correlation between irrigable outdoor space and outdoor water consumption [3]. They hypothesize this is due to inefficient irrigation practices. Syme et al. [34] show that at the household level larger lot sizes increase water consumption. Polebitski et al. [59] find that lot size is an important explanatory factor for seasonal water demand.

3.4 Summary

In this chapter, I described the previous work on identifying, predicting, and explaining outdoor water consumption. Most previous work on identifying outdoor water consumption either requires high-frequency water consumption data or assumes that all winter consumption is for indoor use and all increases in summer consumption are related to outdoor water consumption. Additionally, little previous work focuses on predicting disaggregated outdoor consumption.

In the next chapter, I present a method of estimating outdoor water consumption from hourly smart meter data, based on Cole and Stewart's approach [2].

Chapter 4

Data and Preprocessing

In this section I discuss the datasets used in the rest of the thesis (see Chapters 5, 6, and 7) and the preprocessing done on the data. The main dataset consists of hourly smart meter recordings of water consumption from Abbotsford, British Columbia. Secondary datasets are weather information, the National Household Survey results for Abbotsford, and per-household property assessment information from BCAssessment. I describe all of them below, followed by a brief description of how they were combined.

4.1 Smart Meter Water Consumption Data

The primary dataset used is hourly water consumption measurements from the city of Abbotsford, British Columbia. It also contains billing information with an address for each customer. The water consumption data was preprocessed by Steven Wang (see Acknowledgements) and I briefly describe the preprocessing here.

The initial dataset contained water consumption data for more than 20,000 customers. First, the dataset was limited to single-family residential units. Of the single family residential units, records for 873 customers were removed due to network issues that caused periods of no recorded consumption, followed by a very high measurement representing the aggregate value for the previous hours. There were 8229 single-family residential households remaining. The water consumption measurements span the time period from September 1, 2012 to August 31, 2013. The dataset was not recorded in local (Pacific) time, and therefore was adjusted by 7 hours to match the local time.

Additionally, there are two problems with missing or duplicated data. First, there are 96 hours total missing for each consumer, due to maintenance. Second, there is missing and duplicate data related to Daylight Savings Time. For each customer, there are duplicate records at 1:00am on November 4th 2012, and a missing record at 1:00am on March 10, 2013. The duplicate records were averaged to produce a single new value, which was used as the water consumption for that hour.

In total, there are missing records for 97 hours: 1 hour associated with daylight savings time, and 4 days which are missing due to hardware maintenance. Table 4.1 shows the exact hours missing due to maintenance. Note that the maintenance period beginning on March 9th is also 24 hours long, but there is an additional missing record associated with the time change.

Table 4.1: Time periods of data missing due to maintenance

Start	**End**
2013/Feb/16 17:00	2013/Feb/17 16:00
2013/Mar/9 17:00	2013/Mar/10 17:00
2013/Mar/30 17:00	2013/Mar/31 17:00
2013/Jul/27 18:00	2013/Jul/28 17:00

The missing data were estimated using regression tree models. A model was trained for each customer on the non-missing data. The features for the model were the water consumption measurements for 1 hour previous, 2 hours previous, and the hour exactly a week previous. The trained model was used to estimate and fill in the missing values, using previously predicted values as input in cases where there were contiguous missing values.

4.2 Weather

Seasonal water consumption is weather sensitive [8,36,59,60]. This secondary dataset includes daily measurements of temperature and rainfall for the same period as the water consumption data, September 1, 2012 to August 31, 2013. Abbotsford has a temperate climate, with generally heavy rainfall but less during summer. Table 4.2 shows the temperature and rainfall by month. Note that average winter temperatures are above freezing.

4.3 National Household Survey

The National Household Survey[1] (NHS) includes demographic information collected in an optional addition to the 2011 Census, distributed to a subset of households. It contains demographic information such as income information and average family size. The variables used are shown in Table 4.3.

The NHS data is only publicly available at the census dissemination area level. A dissemination area is a contiguous geographic area consisting of multiple census blocks, typically containing 400 to 700 people. See Section 4.5 for a description of how this was combined with the other data available at the household level.

[1] http://www12.statcan.gc.ca/nhs-enm/2011/dp-pd/prof/index.cfm

Table 4.2: Monthly average temperature and total rainfall in Abbotsford, September 2012–August 2013

Month	**Mean Temperature**	**Total Rainfall**
September 2012	15.5 °C	5.5 mm
October 2012	10.5 °C	306.3 mm
November 2012	6.6 °C	240.2 mm
December 2012	3.3 °C	184.2 mm
January 2013	2.5 °C	152.2 mm
February 2013	5.0 °C	103.4 mm
March 2013	7.1 °C	206.4 mm
April 2013	9.1 °C	157.6 mm
May 2013	13.7 °C	101.4 mm
June 2013	16.2 °C	85.0 mm
July 2013	19.4 °C	1.6 mm
August 2013	19.1 °C	57.0 mm

Table 4.3: Variables from National Household Survey data

Variable	**Description**
income	median household total income in dissemination area ($)
family size	average family size in dissemination area

4.4 Property Assesment Data

Information about property values and household characteristics was provided by BCAssesment[2]. The data provided was from 2012. This information is available at the household level with addresses, and contains property values, lot sizes, and building characteristics such as number of bedrooms, and whether the household has a pool. Table 4.4 shows the variables used from this dataset.

Table 4.4: Variables from property assessment data

Variable	**Description**
lot size	lot size (acres)
bedrooms	number of bedrooms
value	assessed value of house ($)
pool code	true if pool is present

[2]https://www.bcassessment.ca/

4.5 Combining Datasets

Water consumption data was retained both at the individual household level for identifying outdoor consumption and aggregated into small neighbourhoods for prediction. The household-level water consumption data was combined by address with the property assessment data only.

The water consumption data for prediction was also aggregated by census dissemination areas, both to allow it to be linked with the NHS data, and because individual water consumption at the hourly and daily levels is extremely variable [46]. For each customer, the address was converted to a latitude and longitude, which was used in combination with the boundary file for census dissemination areas provided by Statistics Canada to map each household to a dissemination area. After this mapping was obtained, the water consumption data for each dissemination area was averaged over its households, to allow a comparison between differently-sized dissemination areas. Figure 4.1 shows the average summer water consumption by dissemination area.

The water consumption data spans 158 dissemination areas, each containing from 1 to 178 consumers. Because small neighbourhoods retain much of the unpredictability associated with individual consumption, for the prediction task only, I removed all dissemination areas with fewer than 50 single-family households. The resulting set is 77 dissemination areas containing 6789 households.

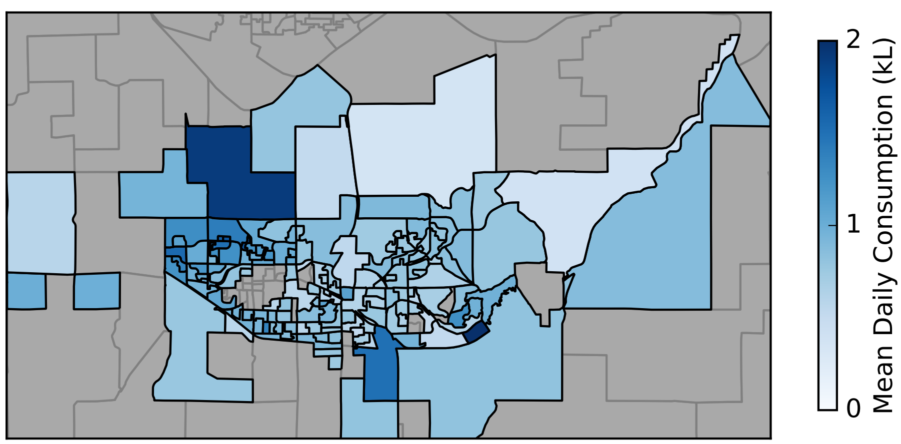

Figure 4.1: Summer water consumption by dissemination area

4.6 Summary

In this chapter, I described the datasets used in the rest of this thesis, and the data cleaning and preprocessing that was performed. In the next chapter, I discuss a method of estimating outdoor water consumption from the hourly household-level data available.

Chapter 5

Identifying Outdoor Water Consumption

As discussed in Chapter 2, identifying outdoor consumption is important for planning and demand management because it consists of high volumes of consumption during relatively short periods, is more easily reducible than indoor consumption, is more variable between households, and contributes proportionally more to peak usage than to average usage. The City of Abbotsford is particularly concerned with identifying and explaining outdoor consumption because of strain on infrastructure during the summer. In this chapter, I describe a method of estimating the amount of water consumption used by a household for outdoor purposes, based on hourly measurements of total water consumption.

Many previous methods of identifying outdoor consumption are not well-suited to disaggregating the data collected from smart meters in practice, which is often recorded at a lower frequency than allowed by the smart meter hardware itself, in order to reduce storage and transmission costs [14]. Although there are disaggregation methods that can accurately identifying individual end uses, they rely on much higher frequency data such as recording consumption every 5s [1], and are therefore less useful when data was not collected specifically for these purposes. On the other hand, older methods that estimate outdoor consumption by subtracting winter use (see Chapter 3) typically use monthly or bimonthly measurements and cannot be used to determine exactly when within a week or at which times of day outdoor water consumption occurs. Additionally, these older methods make the assumption that all seasonal differences in water consumption are due to increased outdoor consumption during the summer. The method of estimating outdoor water consumption described in this section is similar to Cole and Stewart's approach [2], which identifies outdoor consumption from hourly data by setting a maximum plausible threshold for indoor use. This approach relies on the fact that outdoor consumption, such as that used for irrigation and pool filling, is typically high-volume compared to water used for indoor purposes such as bathing, cooking and toilet flushing, and therefore complex disaggregation techniques are less necessary [1]. Additionally, this indoor consumption follows different temporal patterns [8], which can be used to establish a rough threshold.

5.1 Cyclical Patterns in Water Consumption

Water consumption follows three cyclical patterns: seasonal, weekly and daily [7]. The seasonal cycle is associated largely with differences in outdoor water consumption during the summer. I briefly describe these cycles here, because they are used in the next section to contrast the patterns of indoor and outdoor consumption.

The seasonal pattern is associated with greater total water consumption during the summer months, but this pattern may not hold for some end uses, such as showering [35]. Water consumption in the Abbotsford dataset is highest in July and August (see Figure 5.1), despite restrictions on irrigation during these months. This seasonal pattern is likely to be caused by a combination of outdoor consumption and increased indoor consumption due to additional occupancy after the end of the school year.

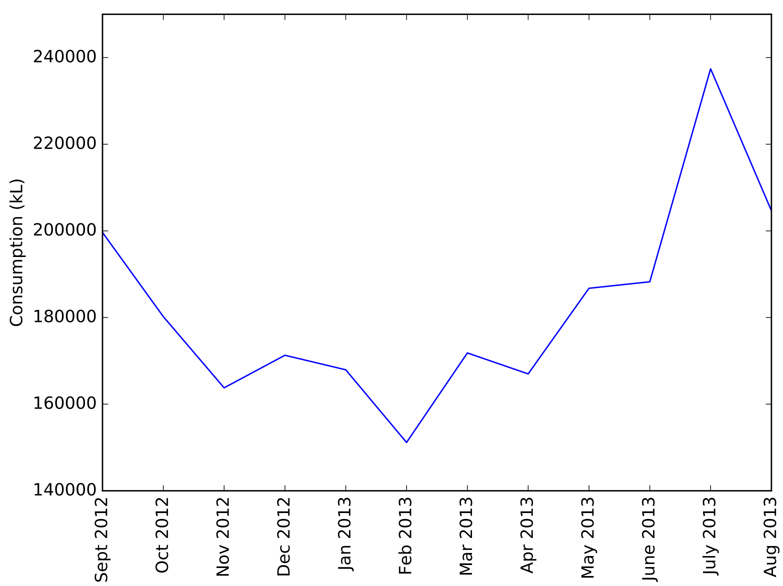

Figure 5.1: Total water consumption for single-family residences by month

There is also a weekly cycle, associated with different consumption patterns on weekends and weekdays. This pattern is much more stable in the winter than in the summer. Figure 5.2 shows a clear weekly pattern with the exception of the week of Christmas (December 25). In contrast, there is no clear weekly pattern in the summer consumption (see Figure 5.3). This is an additional reason why it is useful to analyze outdoor consumption in summer: to determine the reasons for this variability, including whether it is due to outdoor consumption or additional indoor consumption.

Finally, there is a daily cycle, with peaks in the morning and afternoon caused by repeated water consumption habits, such as bathing and toilet flushing [61]. These peaks occur in Abbotsford at 8:00am and 7:00pm (see Figure 5.4).

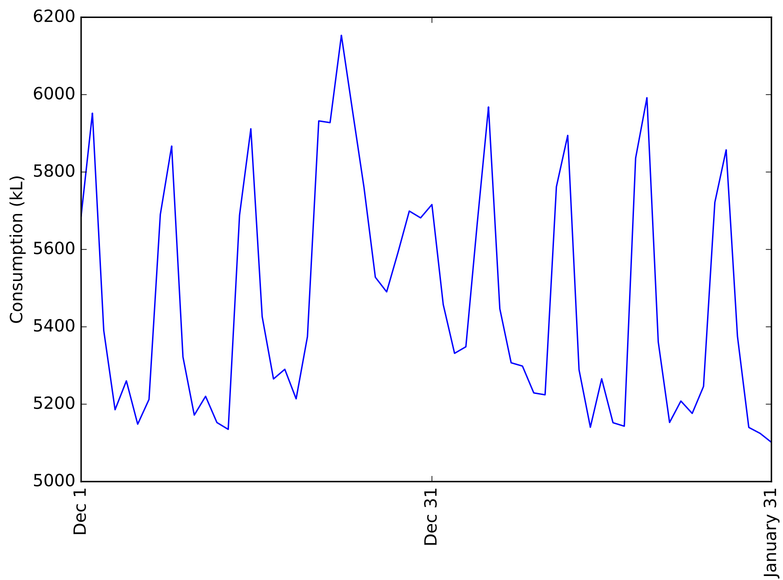

Figure 5.2: Total water consumption for single-family residences in December and January

5.2 Approach

The majority of outdoor water consumption is for two purposes: irrigation and pool filling. Because both of these are very high-volume uses, it is less necessary to use complicated disaggregation techniques requiring high frequency data [1]. Given that outdoor uses require considerably more water in short periods, there should be an hourly consumption threshold that indicates that consumption is unlikely to consist solely of indoor end-uses such as running faucets and flushing toilets. For example, an hourly consumption value in the range of 600 L (litres), more than twice the typical daily per-capita indoor consumption of 262 L [8], is almost certainly used at least partially for outdoor purposes. Therefore, I estimate that all consumption below a particular threshold is used for indoor purposes, and consumption above this range is for outdoor purposes. Formally, after finding a threshold t, I estimate the indoor and outdoor consumption, per-hour and per-household, as:

$$y_{outdoor} = \max{(y_{total} - t, 0)}, \tag{5.1}$$

$$y_{indoor} = \min{(y_{total}, t)}. \tag{5.2}$$

This threshold is established by finding a plausible upper limit on hourly indoor consumption across all households, and analysing when water is used. I validate this by comparing with household characteristics. For example, outdoor consumption should not be as significantly correlated with number of occupants as is indoor consumption.

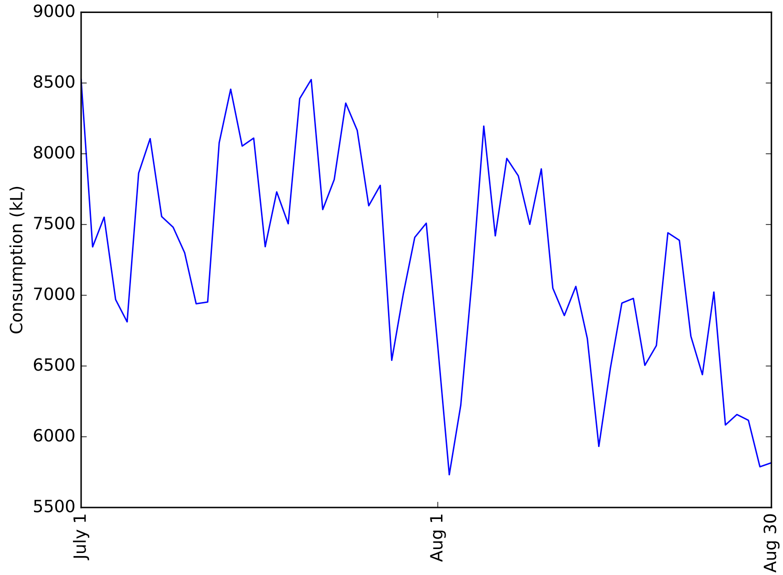

Figure 5.3: Total daily water consumption for single-family residences in July and August

The existence of such a threshold rests on two primary assumptions:

1. Outdoor water consumption is higher-volume than typical indoor usage.
2. Outdoor consumption followings different temporal patterns than indoor use, for example occurring more frequently in the summer.

This method is adapted from Cole and Stewart's [2] approach. Their method also involves selecting a threshold, but if hourly consumption exceeds some threshold then *all* consumption in that hour is counted as outdoor consumption, under the assumption that a significant amount of that consumption is for outdoor purposes. This is sufficient for drawing conclusions about outdoor water consumption on average, but also produces sharp jumps in outdoor consumption around the threshold. (For example, with a 300 L threshold, an hour with 290 L total consumption would contribute nothing to outdoor usage, and an hour with 310 L total consumption would be counted as entirely outdoor usage.) This produces an estimate that does not make sense at the household level, and also would make prediction difficult, which is why I used the approach described previously.

I validate this threshold in large part based on the seasonal patterns of water consumption. At the correct threshold, there should be little outdoor usage for the winter months, although because Abbotsford's climate is temperate, there may still be small amounts of outdoor consumption over the winter. I begin to develop a threshold by looking at plausible indoor uses. I also validate this threshold by showing its relationship to the times when irrigation is allowed in Abbotsford during July and August and how indoor consumption estimated in this way relates to household size.

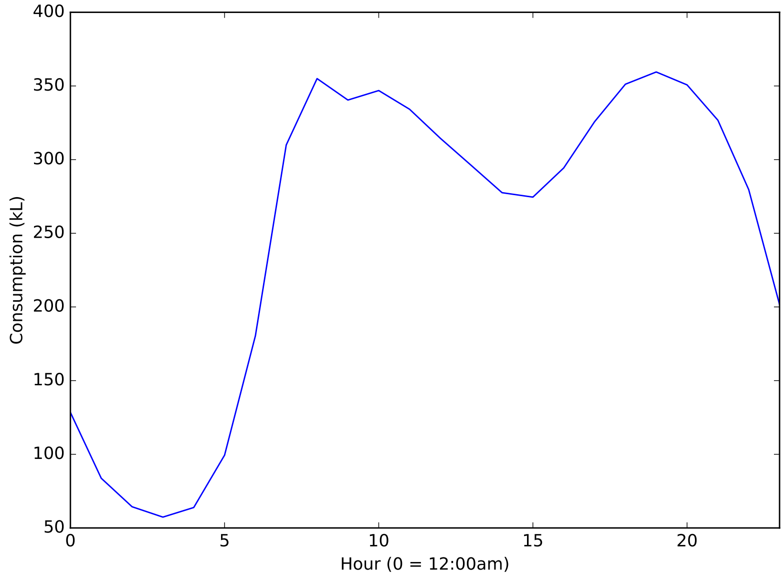

Figure 5.4: Average water consumption per hour over the entire year, summed over all households

5.3 Hourly Consumption Volumes

Total water consumption in the study dataset is 731 L/household/day for single-family units, or about 240 L/person/day. Typical indoor water consumption is 261 L/person/day in North America [8]. Therefore, consumption significantly past this range in a single hour indicates this water is probably used at least partially for high-volume outdoor purposes.

Individual indoor end uses typically require smaller volumes of water than outdoor uses such as irrigation, but they may be combined, especially during the morning and afternoon peaks in consumption. Tables 5.1 and 5.2 show volumes of water used for typical indoor uses, adapted from two versions of the Residential End Uses of Water study in 1999 [8] and 2016 [62]. The shower water consumption includes both typical flow rate and typical duration of a shower. Faucet consumption is given as a flow rate because the volume per use varies by activity. Typical faucet usage was 8.1 minutes/day in both versions of the study. Note that water consumption for dishwashers and clothes washers has significantly decreased between 1999 and 2016; data from both years are included for reference because rates of ownership of more efficient appliances in Abbotsford are unknown.

Figure 5.5 shows the hourly consumption in each range. Consumption is concentrated in the lower ranges: most consumption is the result of many households using small amounts of water per hour, rather than more infrequent but higher-volume consumption. This indicates that, taken over the whole year, most consumption is for (low-volume) indoor purposes which is consistent with Abbotsford's temperate climate.

Table 5.1: Volumes of typical indoor end-uses (1999)

End Use	Volume
Faucet	4.9 L/min
Toilet	13.2 L/flush
Shower	65 L/shower
Clothes Washer	154 L/load
Dishwasher	37 L/load

Table 5.2: Volumes of typical indoor end-uses (2016)

End Use	Volume
Faucet	4.9 L/min
Toilet	9.8 L/flush
Shower	62 L/shower
Clothes Washer	117 L/load
Dishwasher	23 L/load

Figure 5.6 shows the water consumption in larger ranges. These ranges represent, roughly, low-volume indoor uses (<100 L), multiple or higher-volume indoor uses (100 to 300 L), some outdoor consumption (300 to 600 L) and primarily outdoor consumption (>600 L). This is consistent with Table 5.1 and 5.1, which show most individual indoor end-uses require well under 100 L of water. Note that there is only a relatively small percentage of water use identified (over the whole year) as outdoor consumption, which is consistent with previous work [8] that shows this percentage to be around 20% in temperate climates.

5.4 Seasonal Patterns and Developing a Threshold

As in previous work on identifying outdoor consumption without high-frequency measurements (see Chapter 3), I largely rely on making comparisons between summer and winter consumption. The approach taken here differs from most of the previous work in that it produces an hourly estimate of outdoor consumption and does not assume that all differences in summer and winter consumption are the result of increased outdoor consumption during the summer.

Abbotsford's temperate climate means that it is possible for there to be some water use in the winter, but it should be considerably less frequent than during the summer months. In this section I compare the consumption for the months of July and August, when Abbotsford enforces watering restrictions due to potential water shortages, to December and January, which are typically the coldest months in Abbotsford. Figure 5.7 shows the differences in consumption during these periods. Lower-volume usage is relatively consistent between summer and winter, up to the

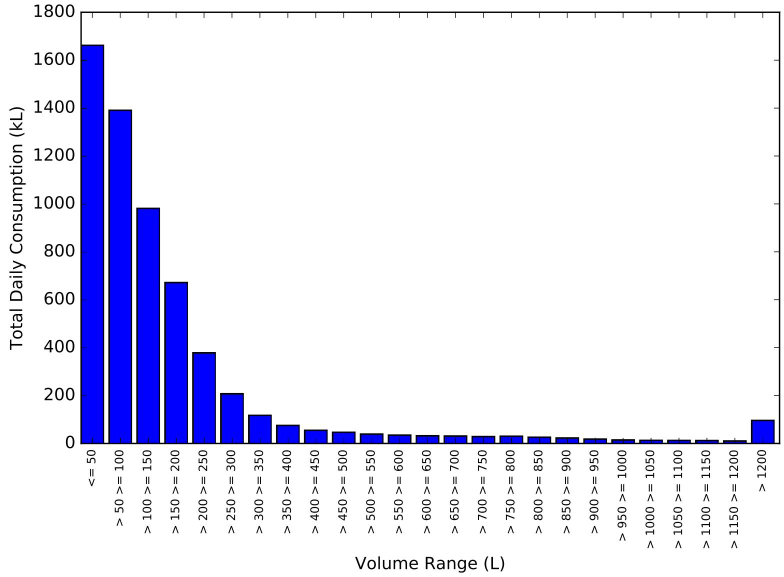

Figure 5.5: Percentage of consumption by hourly volume range

150 L to 200 L volume range. This is consistent with the idea that this is primarily indoor use. I chose a threshold of 300 L/hour as the point past which usage is likely to be for outdoor purposes, because summer consumption may include some extra indoor use due to extra occupancy. There is a tradeoff between recognizing all outdoor use and not miscategorizing somewhat higher-volume indoor use (such as washing machines or many lower-volume uses occuring simultaneously). The 300 L threshold makes this tradeoff conservatively, capturing much outdoor use with little indoor use, although some lower-volume outdoor consumption may be missed. In the 300–350 L/hour range, the ratio between indoor and outdoor volume ranges is 2.4, which suggests this range contains significant amounts of outdoor consumption.

5.5 Sensitivity of the Threshold

In this approach, any disaggregation threshold is inherently approximate. I evaluate thresholds on both sides of 300 L to determine how sensitive the approach is to changes in the threshold. Ideally, a good threshold will result in indoor usage that is relatively constant over the year, and outdoor usage that peaks in summer. Additionally, outdoor usage should not be significantly correlated with the number of household occupants, whereas indoor usage should be significantly correlated with the number of occupants. I evaluate 200 L, 300 L, and 400 L as potential thresholds t for separating indoor and outdoor consumption as described by Equations 5.2 and 5.1. If small changes to the threshold result in large changes to the outcome, this would make the approach untenable given that there is no ground truth with which to evaluate the threshold. However, as described below, there are only small changes, so it is simply a matter of making the tradeoff described previously between capturing all outdoor

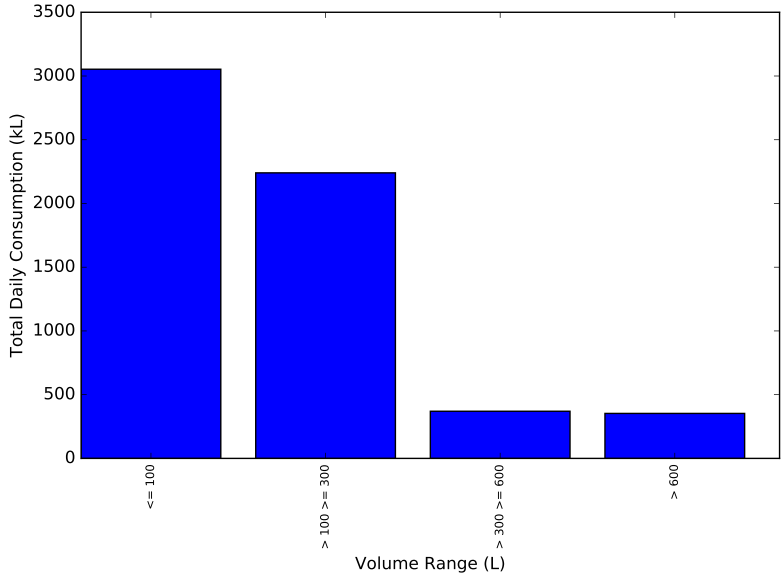

Figure 5.6: Percentage of consumption by hourly volume, large ranges

consumption and capturing only outdoor consumption. I validate the threshold by showing how the indoor and outdoor consumption estimated with a given threshold relates to seasonal patterns, the days when irrigation is permitted, and the number of bedrooms per household.

Small variations in indoor consumption are expected between summer and winter, but indoor consumption should still be relatively constant over the year. Figure 5.8 shows the estimated indoor consumption given each threshold. A higher threshold results in a more pronounced seasonal pattern because more outdoor consumption is misclassified as indoor consumption. Note that for thresholds, the consumption over the year is relatively constant compared to total consumption. Figure 5.9 shows the estimated outdoor consumption at each threshold. For the lowest threshold (200 L), a significant amount of outdoor consumption is identified during the winter months, but both the 300 L and 400 L thresholds show relatively little outdoor consumption during the winter. The 400 L threshold shows the least outdoor consumption in winter, but captures less of the outdoor consumption during the summer.

During the summer of 2013, Abbotsford enacted watering restrictions during July and August. These watering restrictions state that residents can water their lawns using sprinkler systems only between the hours of 6:00 am and 8:00 am on designated days. These watering restrictions prohibit sprinkler use outside of the designated times, but other outdoor water uses are permitted. Even numbered houses are restricted to watering on Wednesdays and Saturdays; odd numbered houses are restricted to watering on Thursdays and Sundays. The figures below show the consumption per-day, separated by odd and even house numbers. At the lowest threshold, significant amounts of water consumption near the 7:00 am peak are identified as outdoor usage, but the increased high-volume water consumption is likely due to

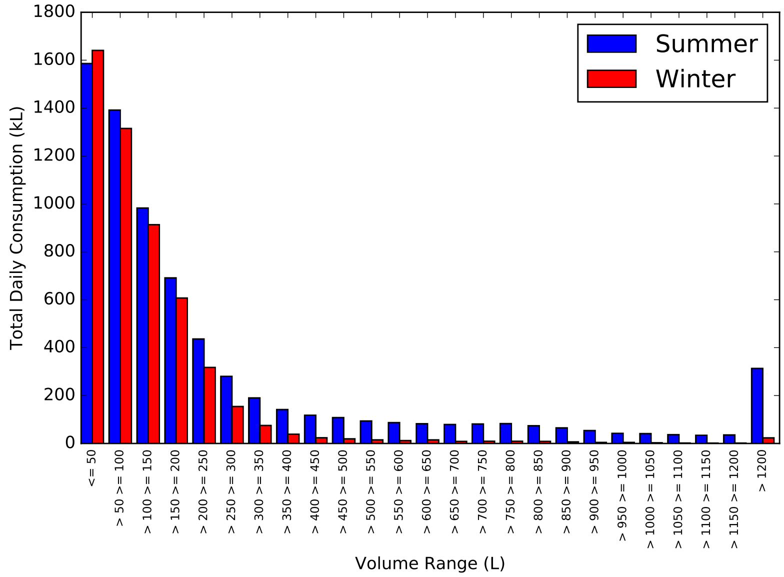

Figure 5.7: Percentage of consumption by hourly volume, summer and winter

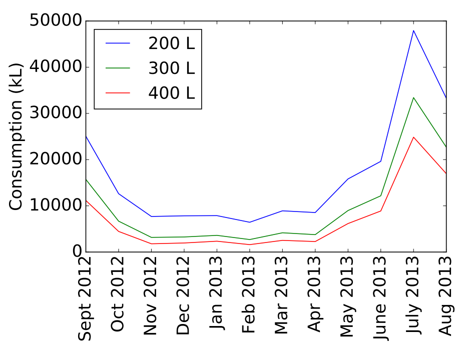

Figure 5.8: Estimated outdoor consumption by month

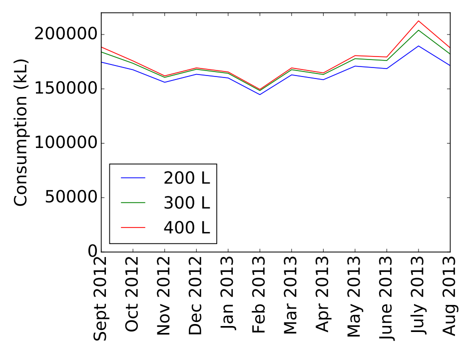

Figure 5.9: Estimated indoor consumption by month

simultaneous indoor end-uses at this time. At the highest threshold, significantly less outdoor consumption is identified. Note that this outdoor consumption peaks slightly later than the peak for total consumption, so it is plausible that there is also some outdoor use at this time, timed to avoid evaporation.

Finally, outdoor consumption should be relatively similar regardless of the number of occupants in the house, because it does not scale in the way that indoor consumption does, although there may be some relationship to outdoor consumption through greater house and lot size. Similarly, indoor consumption should depend on household size. I use the number of bedrooms to estimate occupancy, because average number of people per house is not available at the household level. Figures 5.16–5.18 show the relationship between number of bedrooms and outdoor water consumption at each threshold. The lowest threshold categorizes significant amounts of winter consumption as outdoor use for all households, while the highest threshold identifies less

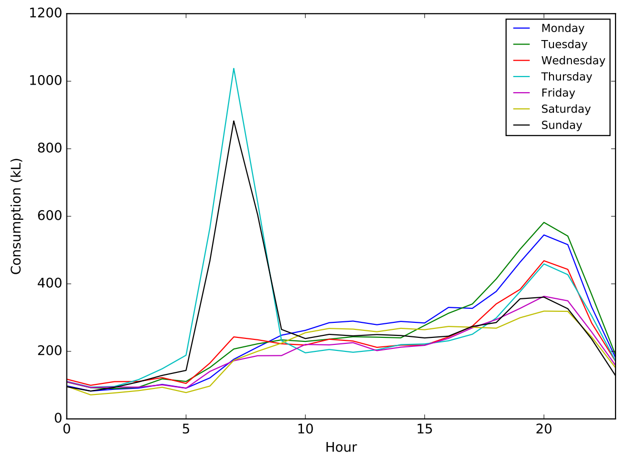

Figure 5.10: Odd-numbered houses, 200 L threshold

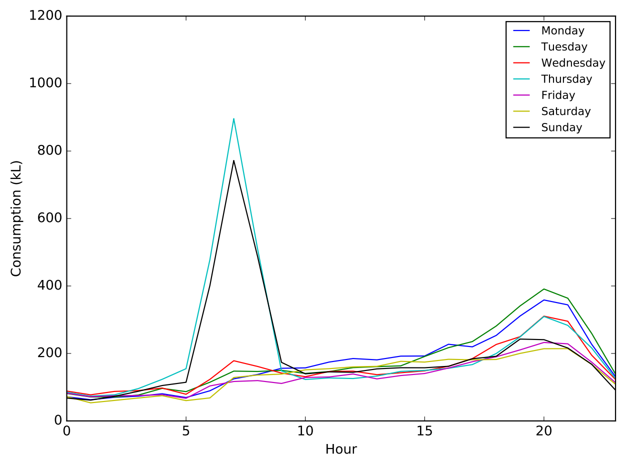

Figure 5.11: Odd-numbered houses, 300 L threshold

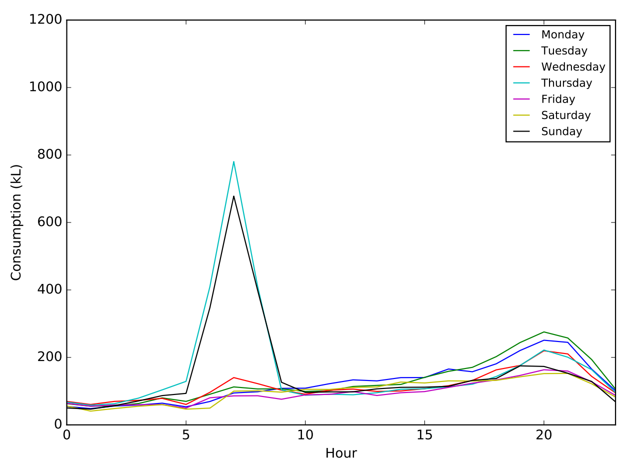

Figure 5.12: Odd-numbered houses, 400 L threshold

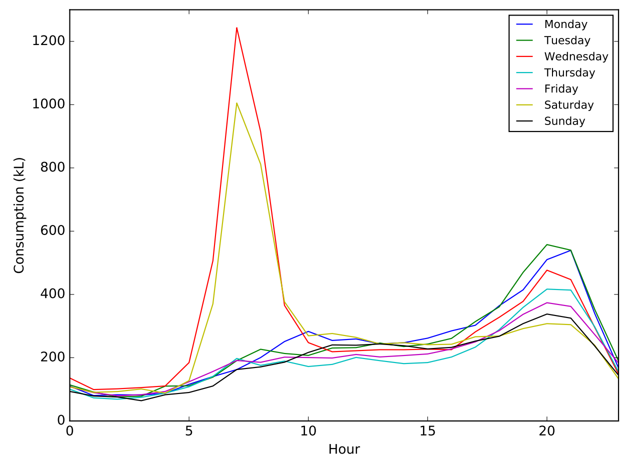

Figure 5.13: Even-numbered houses, 200 L threshold

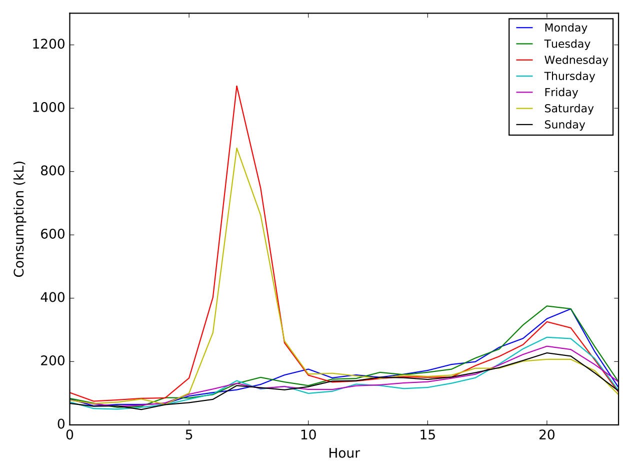

Figure 5.14: Even-numbered houses, 300 L threshold

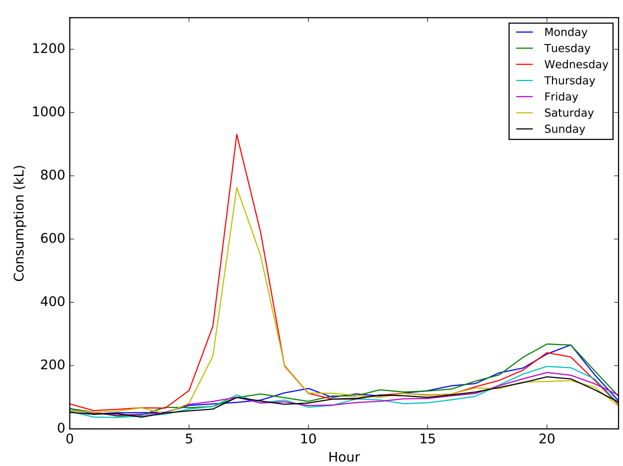

Figure 5.15: Even-numbered houses, 400 L threshold

outdoor consumption in the summer without significantly decreasing the amount of outdoor consumption identified in the winter. At the middle threshold, the outdoor consumption is relatively constant for all numbers of bedrooms, which is consistent with the idea that outdoor consumption should vary less by household size than indoor consumption. Figures 5.19–5.21 show the relationship between number of bedrooms and indoor consumption, which for all household sizes is more consistent over the year than outdoor consumption. Note that on average, households with more bedrooms have higher indoor water consumption.

The previous discussion shows that while the threshold is relatively robust to changes, raising or lowering the threshold from 300 L means potentially miscategorizing more indoor use or not identifying significant amounts of outdoor use. Therefore, I use a 300 L threshold for identifying outdoor water consumption in the rest of the thesis. Based on Equation 5.1, the outdoor consumption at each hour (in litres) is now defined as:

$$y_{outdoor} = \max\left(y_{total} - 300, 0\right). \tag{5.3}$$

5.6 Peak and Average Consumption

While the total amount of outdoor consumption identified is relatively low, this estimation of outdoor water consumption is still useful because outdoor consumption contributes proportionally more to peak consumption than to average consumption. Figure 5.22 shows the indoor and outdoor consumption at the peak hour and on average. Note that indoor consumption is also greater at the peak hour because the average hour includes overnight consumption, which is low in all volume ranges. Figure 5.23 shows the consumption on the peak day. While outdoor consumption is relatively low, even on the peak day, the greater variability in outdoor consumption between the average and peak days means that it is useful to be able to estimate this consumption.

5.7 Summary

In this chapter, I described a method of estimating outdoor water consumption from hourly data by setting an upper threshold on plausible indoor consumption. I showed that this threshold is relatively robust to changes, and chose 300 L/hour as a threshold for the remaining work. I also showed that the contribution of outdoor consumption to total consumption is relatively greater for peak periods of use than it is for average-use periods.

In the next chapter, I take the estimate of outdoor consumption produced using this method and create a model to predict hourly outdoor water consumption.

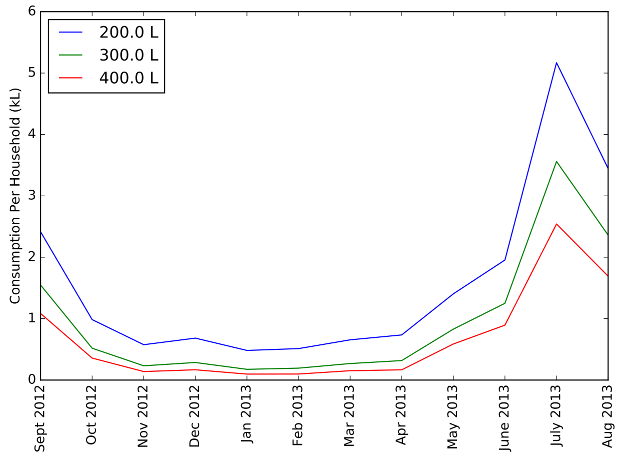

Figure 5.16: Outdoor consumption, 1 and 2 bedroom houses

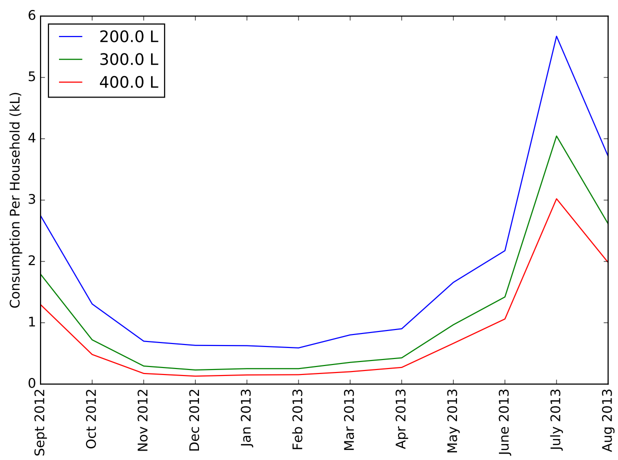

Figure 5.17: Outdoor consumption, 3 bedroom houses

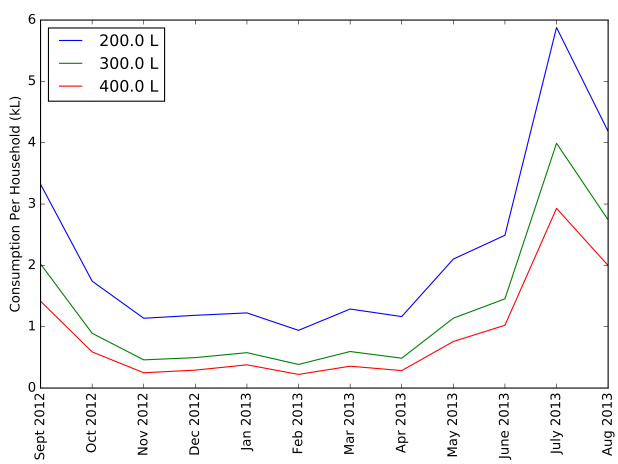

Figure 5.18: Outdoor consumption, 4-plus bedroom houses

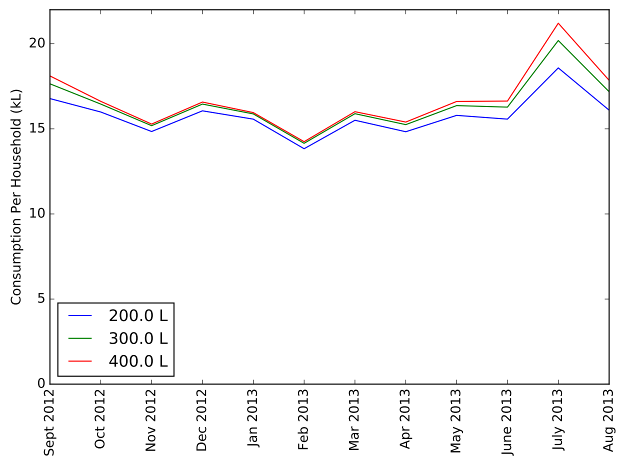

Figure 5.19: Indoor consumption, 1 and 2 bedroom houses

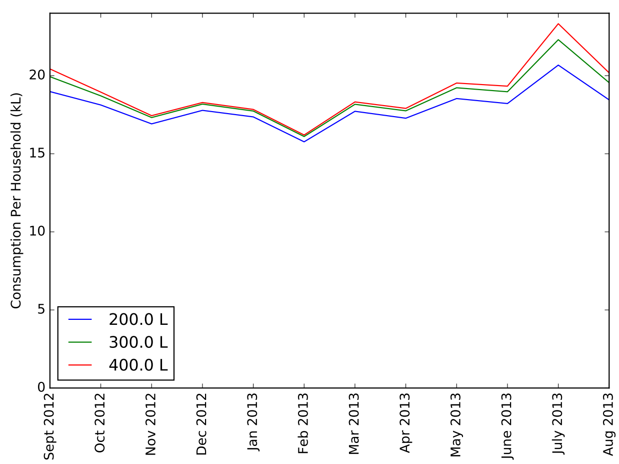

Figure 5.20: Indoor consumption, 3 bedroom houses

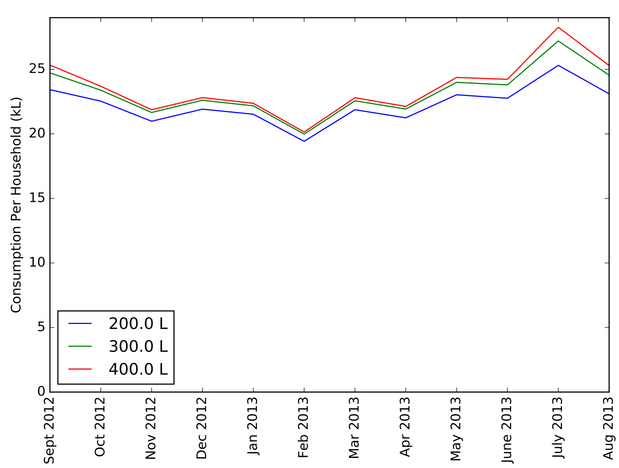

Figure 5.21: Indoor consumption, 4-plus bedroom houses

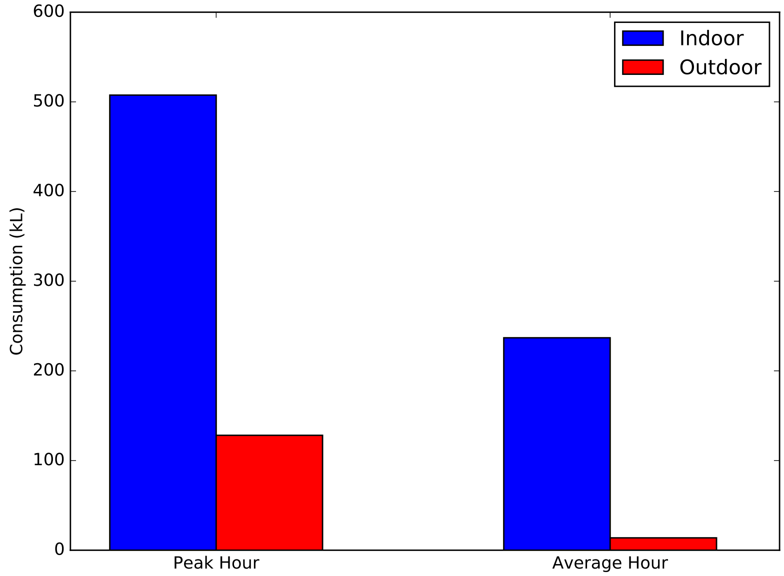

Figure 5.22: Peak and average hour consumption

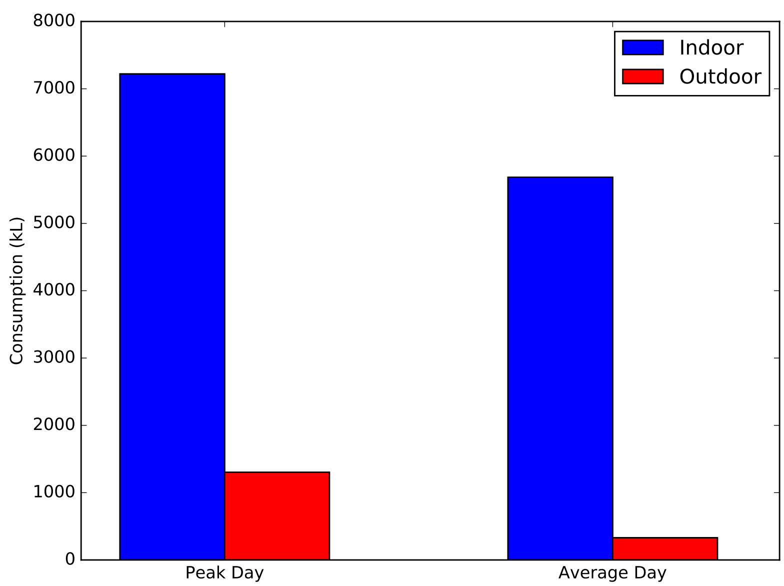

Figure 5.23: Peak and average day consumption

Chapter 6

Predicting Outdoor Water Consumption

In this chapter, I present a model for predicting outdoor hourly water consumption, and discuss the challenges of this prediction task by comparing to a model for prediction of total outdoor consumption. The prediction task described in this chapter differs from previous work in that it attempts to predict disaggregated outdoor rather than total consumption, and additionally predicts over a smaller spatial scale than does most previous work. Previous work has found similar difficulties with predictability for single-household consumption [46].

Accurate prediction of outdoor water consumption would be useful for short-term management in that it could be used to better estimate the effects of variations in weather and of restrictions on short-term water consumption. I focus on hourly prediction because, although daily prediction would be more useful for managing water restrictions, daily consumption appears to be more unpredictable. I use a model based on ensembles of regression trees in order to maintain some interpretability of the models because knowing the determinants of hourly water consumption would be useful for planning. Previous work has found good results with ensembles of trees for predicting total water consumption [44].

6.1 Problem Definition and Model Structure

The prediction task is essentially time series prediction and can be formulated as a supervised learning task (see Chapter 2). Previous measurements of water consumption and additional features are used to predict the next hour's water consumption.

I develop models for predicting outdoor hourly water consumption for individual dissemination areas. The water consumption is normalized by dividing by the number of households in the dissemination area so that the models will not have to take into account the varying sizes of disseminations areas, and only the 77 dissemination areas with greater than 50 single-family residences are included. Only summer consumption

(July and August) is predicted because this is when there is the greatest concern about water demand in Abbotsford.

Formally, the prediction task is:

$$\hat{y}_t = \hat{f}(y_{t-1}, y_{t-2}, y_{t-3}, y_{t-168}, x_1, ..., x_n)$$

Where $\hat{y}$ is the predicted value for next-hour water consumption, f is the model for prediction, $y_{t-1}, y_{t-2}, y_{t-3}, y_{t-168}$ are hourly values of actual water consumption for the previous 3 hours and the same hour the previous week and $x_1, ..., x_n$ are other features given in Table 6.1. Note that the values of y for the models predicting outdoor consumption are estimated using the method described in Chapter 5. Weather and date variables are at a daily frequency and are repeated for each of the relevant 24 hours. Property and demographic features are per dissemination area. Water consumption values for the previous several hours, as well as the same hour one week previous, were found to be good predictors of hourly water consumption in previous work [44, 46].

Table 6.1: Model features

Category	**Feature**	**Description**
weather	rainfall (current)	rainfall amount
	rainfall (last 3)	rainfall amount for previous 3 days
	rain occurrence (current)	true for days when rain occcured
	rain occurrence (last 3)	true when rain occurred over last 3 days
	temperature (current)	daily high temperature
	temperature (previous)	high temperature for previous day
date	weekday	true for weekday
	watering	true for days when sprinklers are permitted
property	lot size	average lot size
	bedrooms	average number of bedrooms
	value	average value of houses
	pools	percentage of houses with pool
demographic	income	median household income
	household size	average household size

The models are gradient-boosted ensembles of regression trees trained using the LSBoost algorithm [19] which is implemented in the MATLAB function LSBoost[1]. (See Chapter 2 for a description of boosted regression trees.)

Because water consumption varies between weekends and weekdays, two outdoor models are developed in order to determine if it improves accuracy to use separate models for predicting weekends and weekdays. In addition, a model is trained to predict total consumption for comparison.

[1]https://www.mathworks.com/help/stats/framework-for-ensemble-learning.html

1. The *Outdoor₁* model predicts outdoor consumption and uses all features given in Table 6.1.

2. The $Outdoor_2$ model consists of two separate ensembles for predicting outdoor water consumption on weekends ($Outdoor_{weekend}$) and weekdays ($Outdoor_{weekdays}$). The *weekday* feature is omitted.

3. The *Total* model predicts total water consumption using all features.

6.2 Model Training and Evaluation

There are two goals in developing models: to improve model accuracy, and to give an unbiased estimate of how the model will perform on unseen data. The models are developed in several steps:

1. 10 dissemination areas are selected randomly and reserved for testing combinations of model parameters, so that the model parameters are not overfit to the set used for evaluating performance.

2. For each set of parameters, combinations of parameters are trained using cross-validation over dissemination areas. For each combination of parameters, 10 models are developed, each with 1 dissemination area left out to evaluate performance. Good parameters are chosen based on performance.

3. Based on the model parameters chosen, models are trained and evaluated using leave-one-out cross-validation [20] over the dissemination areas. This step uses the 67 dissemination areas that were not used to choose the model parameters. On each iteration, a model is trained on 66 of the 67 dissemination areas. On each iteration, the accuracy is calculated using the left-out dissemination area as testing data. These results are discussed in Section 6.4.

Absolute errors are given in the results section for each model. The absolute error is calculated as:

$$|y_t - \hat{y}_t|.$$

The mean absolute error (MAE) is the average of the absolute errors:

$$\frac{1}{n}\sum_{i=1}^{n} |y_i - \hat{y}_i|.$$

6.3 Parameter Selection

Based on the reserved set of 10 dissemination areas discussed previously, I chose values for two parameters that are likely to produce good performance on the remaining dissemination areas. I optimized over two parameters: maximum number of splits per tree, and the number of trees per ensemble. The maximum number of splits is the total number of non-leaf nodes in the tree. Figure 6.1 shows the accuracy over these parameters, for the *Outdoor₁* model. While the strictly optimal value for the maximum number of splits is 2, the accuracy is relatively sensitive to the number of trees in the ensemble. I instead chose 1 split and 80 trees per ensemble because the performance is not significantly affected by the ensemble size past that point, suggesting those parameters will result in more stable performance for the final models. While it would have been ideal to perform parameter selection within the main cross-validation step and choose the optimal value on each iteration, it was prohibitively slow.

I chose the same parameters for all models. The maximum number of splits is also not strictly optimal for the other models, however it was similarly close-to-optimal and chosen because the accuracy was less sensitive to variation in the ensemble size and therefore likely to transfer better to the remaining data.

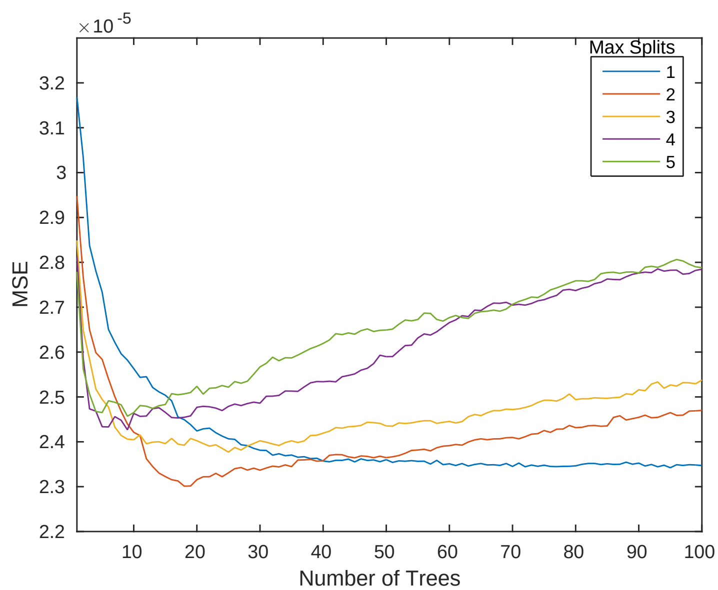

Figure 6.1: MSE for number of trees and maximum number of splits used as model parameters

6.4 Results

In this section I compare results for all models, and although the predictive accuracy is not high, the comparison with total consumption gives insights into why this is the case.

While the outdoor models show a general ability to predict the direction of consumption, they do not perform well at predicting the magnitude of consumption, especially for high values. Previous work with predicting individual consumption had similar difficulties with predicting the magnitude of peaks [46]. Figure 6.2 shows the predicted outdoor consumption for the largest dissemination area. Note that the prediction of peaks is often one sample late, because as discussed in the next section, the model predicts largely based on the previous hour's consumption. Figure 6.3 shows the predicted total consumption for the same dissemination area. Although overall performance is better for total consumption, the model still fails to predict the magnitude of the morning and afternoon consumption peaks accurately, and sharp peaks in consumption are typically not predicted accurately.

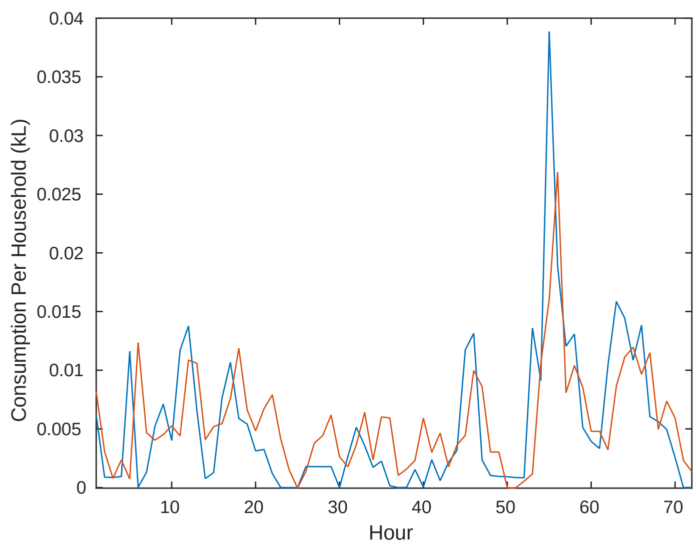

Figure 6.2: Prediction of *Outdoor₁* compared to actual value for the largest dissemination area in the first 3 days of July. The orange line is the predicted consumption value and the blue line is the actual value.

Tables 6.2–6.4 show the errors in kilolitres for each model type. The tables should be interpreted in the light of the fact that the model predicts *mean* water consumption per household in a dissemination area. Although the model errors are small in absolute terms, they become more significant in terms of the average hourly values of 0.0045 kL (4.5 L) for outdoor water consumption and 0.0353 kL (35.3 L) for total water consumption. Note that for both types of consumption, the average errors are significantly higher than the median, suggesting that the poor performance is in large part due to large errors rather than many small errors.

The overall error for both the $Outdoor_1$ and $Outdoor_2$ models are similar, despite performance being significantly worse on the weekends. A potential reason that separate models are not useful is discussed in the next section.

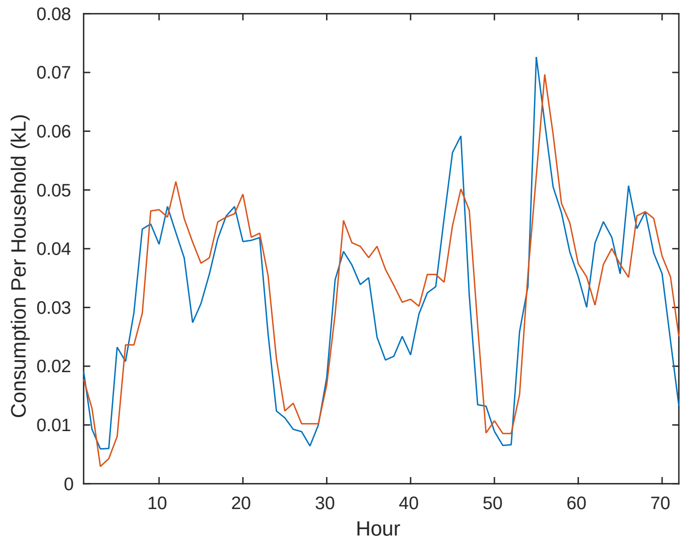

Figure 6.3: Prediction of *Total* compared to actual value for the largest dissemination area in the first 3 days of July. The orange line is the predicted consumption value and the blue line is the actual value.

Weekends and watering days when sprinkling is permitted have high errors for all models. This suggests that this variability is hard to predict.

Table 6.2: Absolute errors for $Outdoor_1$ (kL)

	average	**median**	**75th percentile**	**90th percentile**
overall	5.054×10^{-3}	2.819×10^{-3}	6.453×10^{-3}	1.212×10^{-2}
weekdays	4.877×10^{-3}	2.752×10^{-3}	6.281×10^{-3}	1.178×10^{-2}
weekends	5.558×10^{-3}	3.027×10^{-3}	7.010×10^{-3}	1.330×10^{-2}
watering	5.407×10^{-3}	3.040×10^{-3}	6.890×10^{-3}	1.281×10^{-2}
no watering	4.612×10^{-3}	2.572×10^{-3}	5.938×10^{-3}	1.134×10^{-2}

The prediction accuracy varies significantly by hour. Figures 6.4–6.7 show the MAE by hour. Note that in all cases the error is highest at the morning peak, which probably is related to more variability in consumption during these times. In addition, even predicting total water consumption for the summer is more difficult than it would be for winter consumption due to higher variability, which explains the relatively low accuracy.

6.5 Interpretability

Ensembles of regression trees were selected as a model because of their *relative* interpretability compared to, for example, neural network models. However, ensembles

Table 6.3: Absolute errors for *Outdoor₂* (kL)

	average	median	75th percentile	90th percentile
overall	5.050×10^{-3}	2.795×10^{-3}	6.431×10^{-3}	1.212×10^{-2}
weekdays	4.867×10^{-3}	2.741×10^{-3}	6.222×10^{-3}	1.185×10^{-2}
weekends	5.570×10^{-3}	2.998×10^{-3}	6.966×10^{-3}	1.301×10^{-2}
watering	5.406×10^{-3}	3.000×10^{-3}	6.868×10^{-3}	1.277×10^{-2}
no watering	4.603×10^{-3}	2.557×10^{-3}	5.874×10^{-3}	1.136×10^{-2}

Table 6.4: Absolute errors for *Total* (kL)

	average	median	75th percentile	90th percentile
overall	1.172×10^{-2}	8.648×10^{-3}	1.626×10^{-2}	2.584×10^{-2}
weekdays	1.150×10^{-2}	8.497×10^{-3}	1.600×10^{-2}	2.536×10^{-2}
weekends	1.238×10^{-2}	9.074×10^{-3}	1.706×10^{-2}	2.726×10^{-2}
watering	1.211×10^{-2}	8.871×10^{-3}	1.672×10^{-2}	2.667×10^{-2}
no watering	1.124×10^{-2}	8.377×10^{-3}	1.565×10^{-2}	2.486×10^{-2}

of trees lose the easy visual interpretability of single decision tree models. Instead, they can be interpreted using the variable importance, which sums the change in mean-square error for each predictor summed over the number of splits in the model, divided by the number of trees [19]. The predictorImportance feature in MATLAB implements this functionality [2].

In this section, I discuss the variable importances for the models used for predicting consumption for the largest dissemination area. Figures 6.8–6.10 show the variable importances for each model. As is typical [19], they are scaled such that the maximum variable importance is equal to one. In every case, the previous hour's consumption, y_{t-1}, is the most relevant variable. The second most relevant variable for every model is y_{t-168}, with a relative importance of 0.3240 for *Outdoor₁*, 0.2728 for $Outdoor_{Weekday}$, 0.5172 for $Outdoor_{Weekend}$ and 0.1670 for *Total*. Both y_{t-1} and y_{t-168} are excluded from the figures in order to better show the relative importances of the other variables.

Note that while some explanation is possible, the relative importance of correlated variables may be understated based on which was selected first in the model. For example, the differences in the importance of y_{t-2} and y_{t-3} between different models may not be significant. However, the variable importances can still give insights into model performance.

The importance of the previous hour and previous week water consumption in determining water consumption also provides a potential explanation for why modelling weekends and weekdays separately does not significantly help the predictive accuracy of the model. Because the strongest determinants of water consumption, y_{t-1} and y_{t-168}, are the same for each model, having separate models only has the effect of reducing the amount of training data available.

[2] https://www.mathworks.com/help/stats/compactregressionensemble.predictorimportance.html

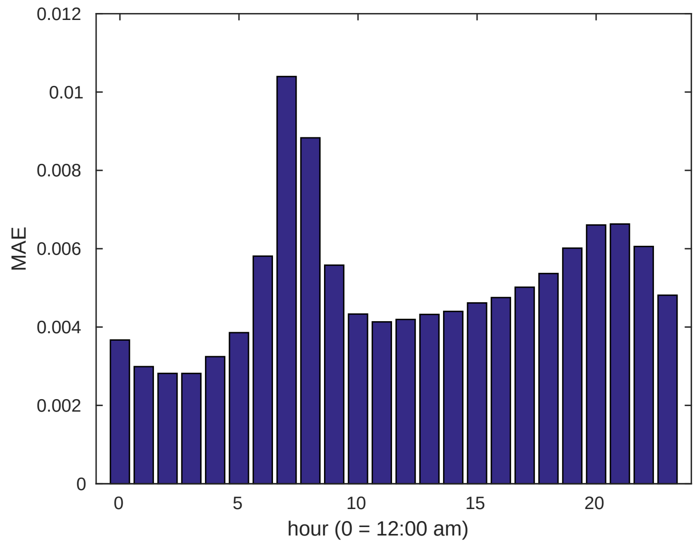

Figure 6.4: MAE per hour for $Outdoor_1$

While the importance of other variables is very small compared to the previous values of water consumption, some variables show up consistently for all models. Weather-related features appear in all models, and although their importance is very minor compared to y_{t-1} and y_{t-168}, an influence from both temperature and rainfall is consistent with previous work on total water consumption [8, 36, 59, 60]. House value appears in the explanatory variables for all models, and median income appears for $Outdoor_{Weekend}$ and $Outdoor_{Weekday}$. Although previous work has focused on the influence of income on both total water consumption [63] and seasonal water consumption [59], the inclusion of house value in all models is consistent with the idea that income has only an indirect influence through factors such as larger house sizes and greater ownership of water-using appliances and pools [57]. Household size or number of bedrooms also appear as slightly important in all models, although previous work has found household size is only strongly correlated with indoor water consumption [8].

6.6 Discussion

It is not clear that the accuracy of the models presented are sufficiently high to be useful in practice for day-to-day management of water systems, or that the models are sufficiently accurate to justify collecting the amount of additional data required beyond previous water consumption values. The models discussed require demographic data and property data, as well as previous temperature and rainfall information and weather forecasts. While the model does provide some ability to forecast next-hour outdoor consumption, the accuracy of any method has to be balanced against the resources required to create a forecast, including collecting the required data [6], and

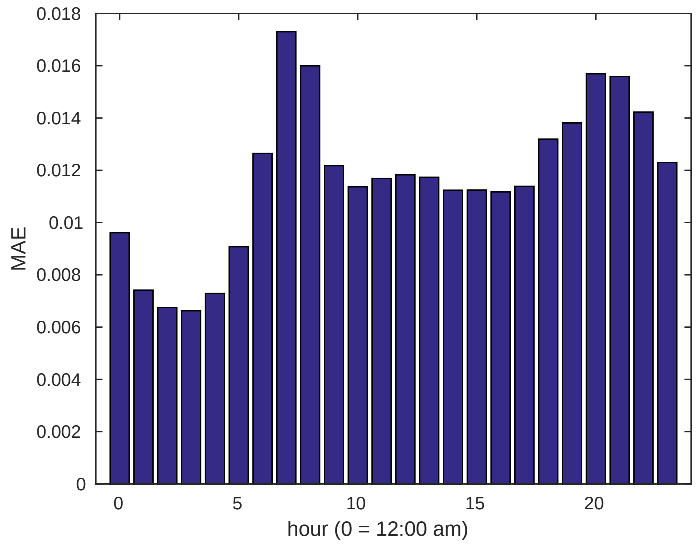

Figure 6.5: MAE per hour for *Total*

particularly for short-term forecasts used for operational purposes, the additional data required creates practical difficulties [9]. The regression tree models were chosen in order to allow interpretability, but the only features that are highly important in reducing model error are previous values of water consumption. It may be the case that accuracy could be improved by adding additional features or by choosing a different machine learning method, but to be practically useful the benefits of the modelling method would need to be balanced against the resources required for the forecast.

Additionally, it would be more useful to be able to predict outdoor water consumption further in advance, such as 24 hours ahead, or to be able to predict daily consumption (ideally several days ahead). This would be useful for making changes to watering restrictions [2]. I found in preliminary work that daily consumption is very difficult to predict as compared to hourly consumption, and this seems to be related to the fact that previous-day and current-day consumption are less correlated than previous-hour and current-hour consumption.

If more data were available, it would also be informative to further aggregate the data in order to determine how large neighbourhoods must be for outdoor water consumption to be accurately predicted. However, given the number of single-family residences in the study and the single-year timespan for the water consumption measurements, significant aggregation would not leave enough data for training a model. There also appears to be significant individual variability in outdoor water consumption (see Chapter 7), which likely constrains predictability except for very large neighbourhoods because the average outdoor water consumption for a dissemination area can be significantly affected by a small number of consumers. Previous work has found that even for entire cities, population size influences the accuracy of predic-

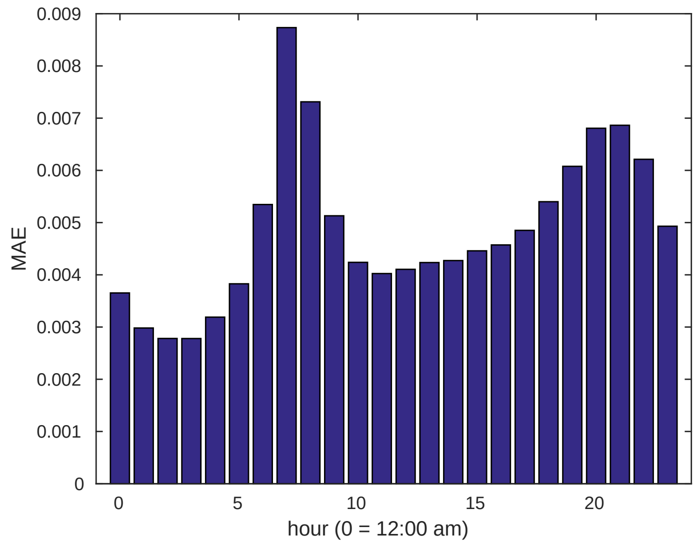

Figure 6.6: MAE per hour for $Outdoor_{weekday}$

tion for total water consumption [43], and the difficulty of predicting outdoor water consumption for small neighbourhoods is consistent with that observation.

6.7 Summary

In this chapter, I presented a model for predicting outdoor water consumption, however it is not clear if the model's accuracy is sufficiently high to be useful for day-to-day management of water systems. I also discussed some of the predictive factors in the models developed.

In the next section, I discuss the explanatory factors for outdoor water consumption over the entire summer.

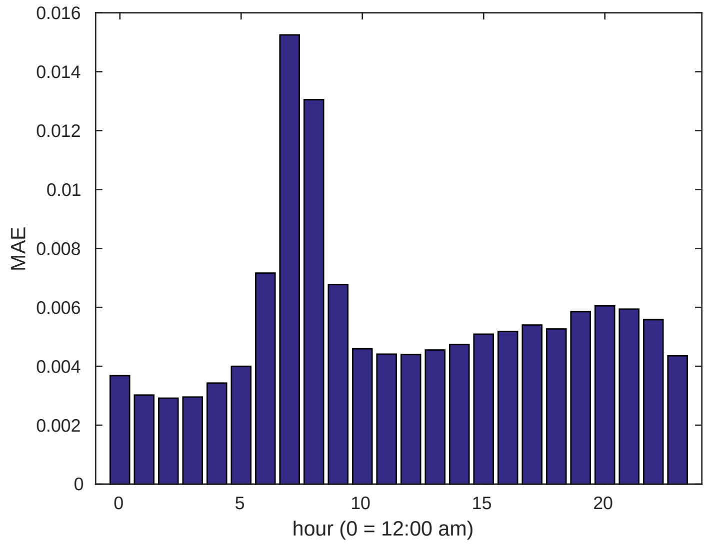

Figure 6.7: MAE per hour for $Outdoor_{weekend}$

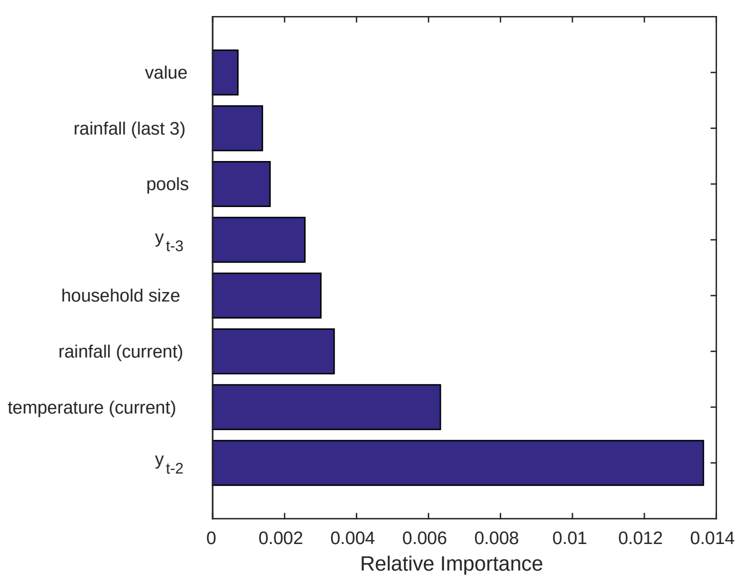

Figure 6.8: Relative variable importance for $Outdoor_1$. The values for y_{t-1} and y_{t-168} are not shown.

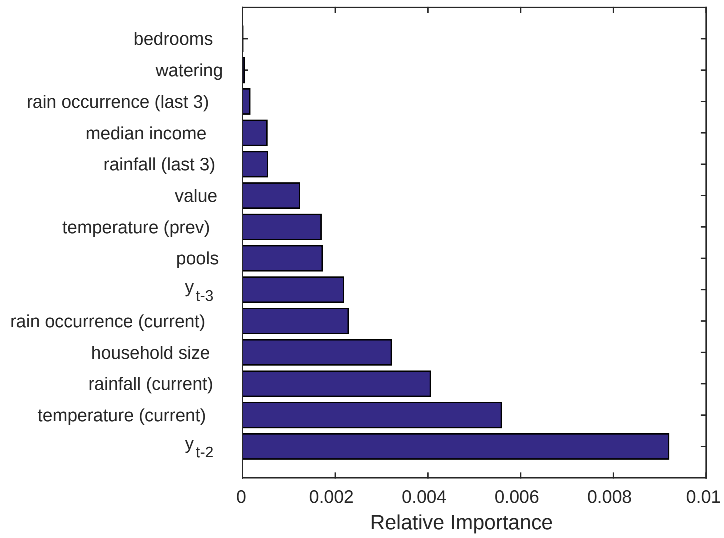

Figure 6.9: Relative variable importance for $Outdoor_{weekday}$. The values for y_{t-1} and y_{t-168} are not shown.

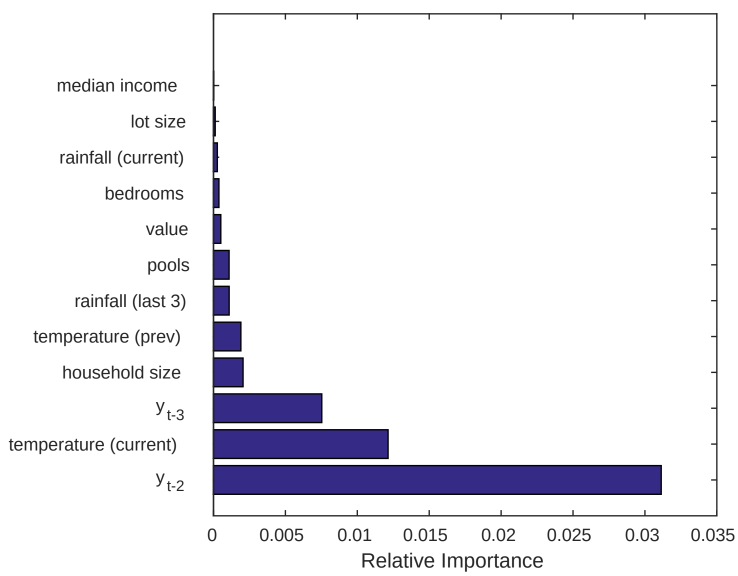

Figure 6.10: Relative variable importance for $Outdoor_{weekend}$. The values for y_{t-1} and y_{t-168} are not shown.

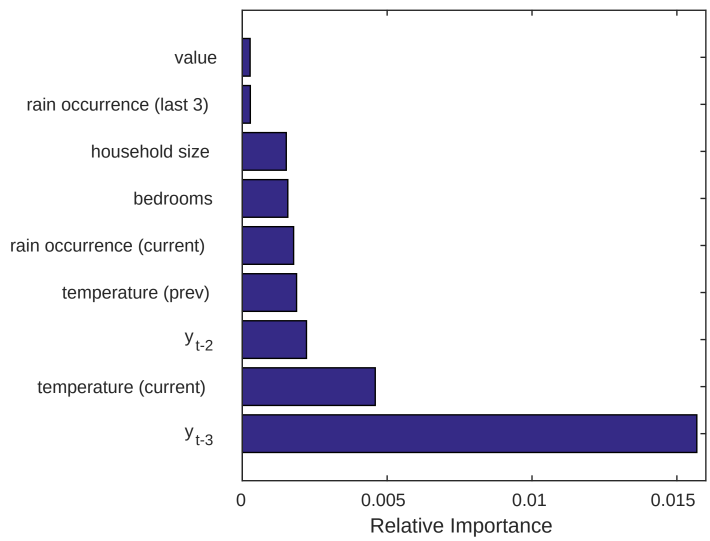

Figure 6.11: Relative variable importance for *Total*. The values for y_{t-1} and y_{t-168} are not shown.

Chapter 7

Explaining Outdoor Water Consumption

The City of Abbotsford is particularly interested in the determinants of outdoor water consumption. Explainable variation in outdoor water consumption between households or between neighbourhoods can be used for demand management and planning watering restrictions [2]. In this chapter, I explain how outdoor water consumption (estimated as described in Chapter 5) during the study period in Abbotsford is related to weather, household income, lot size, and pool ownership. I also show that there is a large amount of variability between households, with a relatively small percentage of households consuming the majority of water used outdoors.

7.1 Statistical Methods

In the following analysis, linear regression is used to model the relationship between outdoor water consumption and demographic and property variables. The correlation coefficient (r) shows the strength of the linear relationship between variables. Interpreting the correlation coefficient requires several assumptions about the underlying data: the residuals (model errors) should have zero mean and constant variance, and be normally distributed and independent [64]. It was determined by visual observation of the data that these assumptions do not hold for all of the data in this chapter; the correlation coefficients from linear regression are given regardless for comparison with previous work, and it is noted where they may not be reliable. To strengthen the results, Spearman's rank correlation coefficient (ρ) is also used to show the strength of a monotonic (rather than linear) relationship between two variables and does not require the same assumptions.

7.2 Temperature and Rainfall

Previous work has shown that water consumption depends on temperature and rainfall (see Chapter 3), and that the sensitivity of water consumption to variation in weather is not consistent across the year but is highest during the summer [36,59,60]. The analysis in this section is restricted to the summer because only a single year of data is available, so it is hard to otherwise separate the effects of changes in weather from unrelated seasonal effects over the year. Further, I only consider the highest-usage months (July and August) because the watering restrictions in these months make it difficult to directly compare the effects of temperature and rainfall relative to the other summer months.

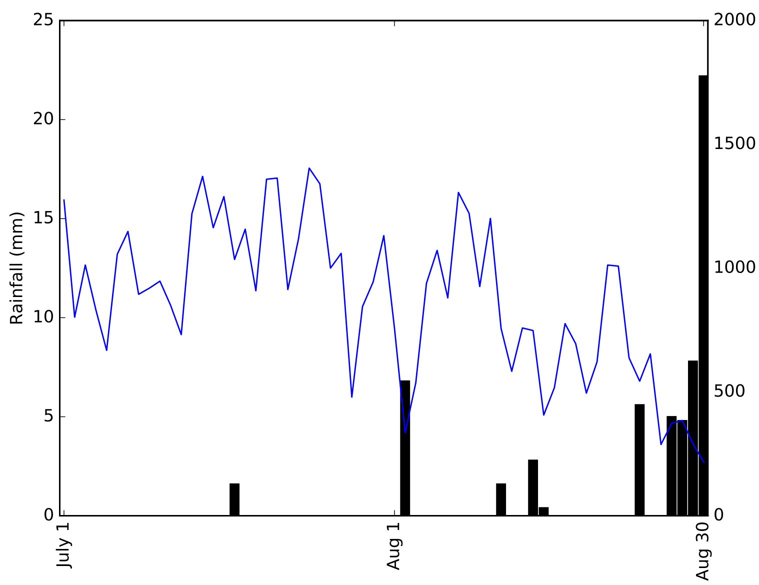

Figure 7.1: Rainfall and outdoor water consumption for all households in July and August 2013

Because a major source of outdoor water demand is irrigation, it is expected that outdoor water consumption will decrease with rainfall. There are clear decreases in outdoor consumption when rainfall occurs, as shown in Figure 7.1. Despite similar average temperatures (19.4 °C and 19.1 °C), the total outdoor consumption for July 2013 was 1037 kL. compared to 698 kL a day in August 2013. This can be explained by the significantly higher and more frequent rainfall in August (57 mm over 9 days) as compared to in July (1.6 mm on a single day). Although total rainfall in both months is relatively low, the drop in outdoor consumption during August is consistent with previous work that suggests the occurence of rainfall (at a weekly timescale) has more effect on water consumption than the amount of rainfall [41].

There is a weak positive correlation ($\rho = 0.401, r = 0.409$) between temperature and outdoor water consumption during July and August. Figure 7.2 shows the temperature and total water consumption for July and August 2013, and Figure 7.3

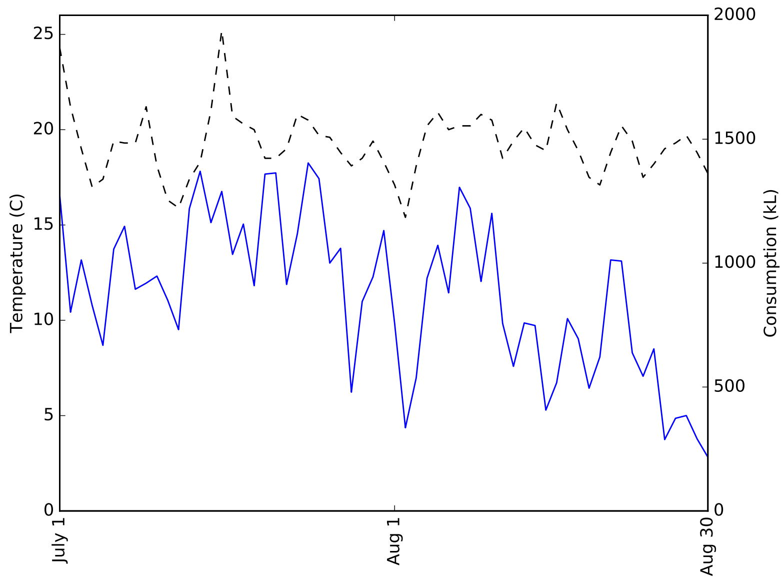

Figure 7.2: Daily high temperature and outdoor water consumption for all households in July and August 2013

shows the correlation between temperature and outdoor water consumption. Correlation with temperature for water consumption has varied between climates in previous work. A study in Portland showed greater correlation with temperature than with rainfall [60], and a study in Pheonix showed greater correlation with rainfall [37]. The low correlation with temperature in Abbotsford is expected because the temperate climate means that water consumption will respond less to evaporation, and because outdoor temperatures were not highly variable during the period analyzed.

7.3 Variability Between Households

Previous work has shown that outdoor water consumption is significantly more variable between households than is indoor consumption [8]. In the study dataset, a small percentage of households account for the majority of outdoor water consumption, while many households use insignificant amounts of outdoor water.

Figure 7.4 shows the distribution of outdoor water consumption (over the entire year) per household. The diagram is restricted to the houses with less than 100 kL of outdoor consumption (98.5 % of households) in order to show patterns at the smaller consumption levels typical of most households. Most households used relatively small amounts of outdoor water and the median consumption is 4.84 kL. Additionally, many households use almost no outdoor water: 1783 households used less than 1 kL of water over the entire year. There are 126 households (1.53%) not shown in the figure which used more than 100 kL of outdoor water, and 31 households (0.38%) which used more than 200 kL of outdoor water. The maximum outdoor water consumption for a single

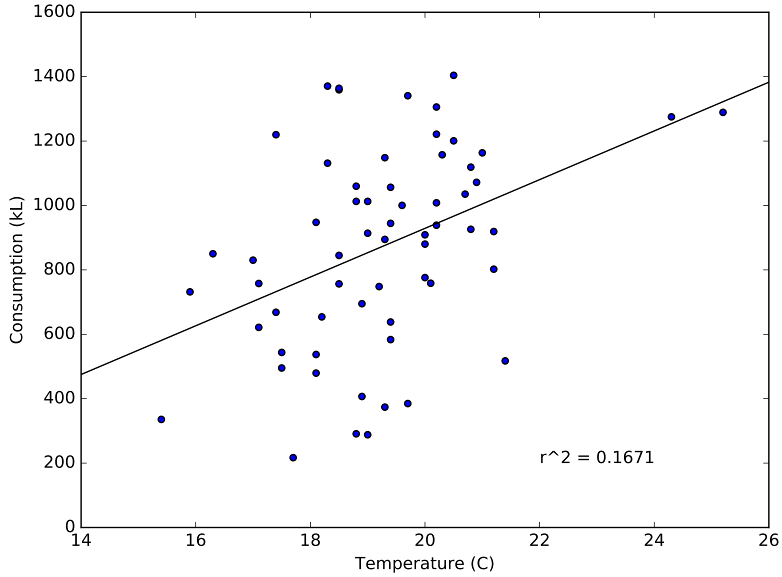

Figure 7.3: Correlation between temperature and daily outdoor water consumption in July and August 2013

household is 1024 kL, about 210 times the median consumption.

In comparison, indoor water consumption is much less variable between households. Figure 7.5 shows the indoor consumption for all single-family households. The maximum indoor consumption for a household is only 7.9 times the median consumption. Similarly, the ratio between the median and the 95th percentile value is 2.4 compared to 10.9 for outdoor consumption. The relatively low variability of indoor consumption is consistent with the idea that indoor usage is primarily determined by the number of people in the household. Even with some additional variation in indoor usage, such as due to water-conserving behaviours or appliance stock, less between-household variation is expected than with outdoor consumption [8].

By ordering the households by their total consumption, it is possible to show how

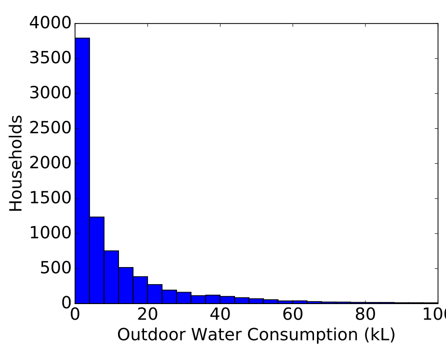

Figure 7.4: Histogram of outdoor water consumption

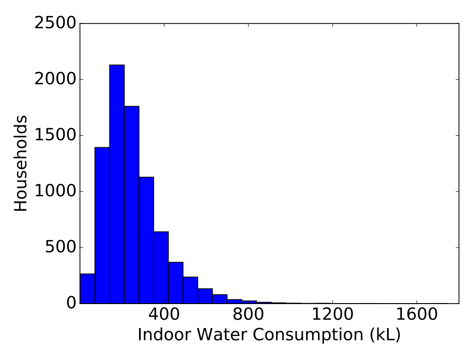

Figure 7.5: Histogram of indoor water consumption

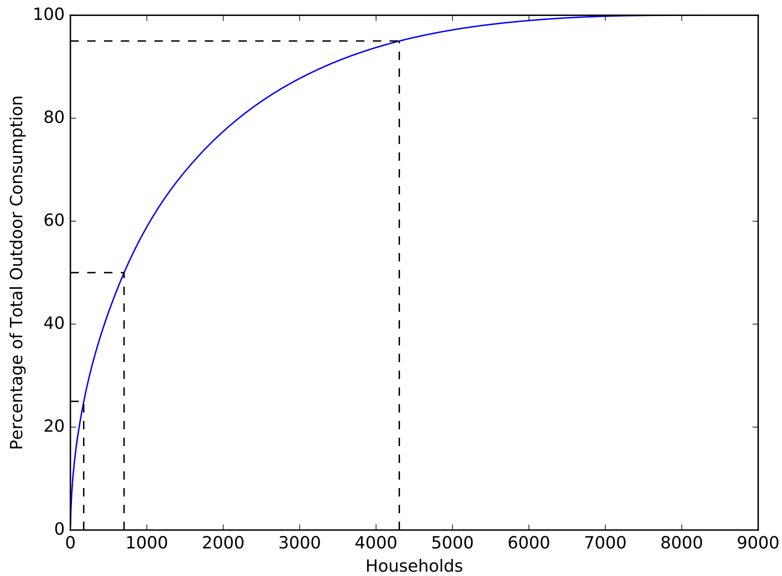

Figure 7.6: Cumulative consumption for first n households, ordered by amount of outdoor water consumption and showing the number of consumers responsible for 25%, 50% and 95% of consumption

much water consumption is explained by the higher-consumption households. Figure 7.6 shows the cumulative consumption for the first n households ordered by outdoor consumption. 25% of outdoor water is used by the 174 households with the highest consumption. 704 households (8.6%) are responsible for 50% of the outdoor consumption, and just over half of households (4305) consume 95% of the water used for outdoor purposes, with the remaining households using almost insignificant amounts of outdoor water. The contribution of the highest-consumption households to total outdoor usage is an important consideration for demand management and also partially explains the difficulty of predicting outdoor consumption for small neighbourhoods described in Chapter 6.

The 3 households that consumed the most outdoor water were checked manually using both the property information and Google Maps[1] to confirm that they are residential properties and that their consumption seems consistent with the property. Only one of these households has a pool or a particularly large lot size, but all have some irrigable landscaping. This is consistent with the following analysis, which shows only a weak correlation between lot size and outdoor water consumption. The water consumption patterns of these households are consistent with using large volumes of water for irrigation.

[1] maps.google.com

7.4 Explanatory Factors at the Dissemination Area Level

Previous work has shown that factors such as income [59] and lot size [38] affect seasonal water consumption at a neighbourhood level. I analyse the relationship between outdoor water consumption at the census dissemination area level (see Chapter 4) and income, lot size, and rates of pool ownership. This analysis is restricted to dissemination areas with greater than 50 single-family residences because the variability in individual consumption described previously would otherwise bias results for areas with few households.

In general, household income correlates positively with water consumption but typically the effects are small [63]. However, this correlation may not be a direct effect of income, but rather caused by differences in lifestyle such as greater ownership of water-using appliances and pools [57], and in that case less variation would be expected when incomes are sufficiently high that lifestyle factors are more homogeneous. In the water consumption data for Abbotsford, there is essentially no correlation ($\rho = 0.122, r = 0.0939$) between income and outdoor water consumption for single-family residential units, as shown in figure 7.7. This may be due to limitations in the demographic data available: the median income is for the entire dissemination area rather than single-family residences.

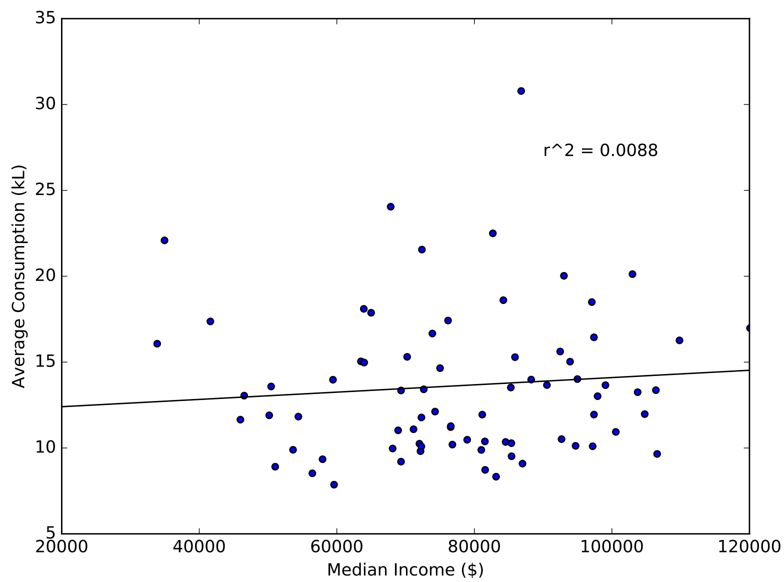

Figure 7.7: Correlation between income and outdoor water consumption by dissemination area

Previous work shows a correlation with between total water consumption and lot size [65]. There is a weak correlation ($\rho = 0.0745, r = 0.423$), shown in figure 7.8, between lot size at the dissemination area level and outdoor water consumption. (The linear regression model overestimates the correlation due to the presence of outliers.)

This may be due to the fact that not all households use outdoor water as shown previously. Additionally, with a few exceptions, lot sizes between dissemination areas are relatively consistent, and the amount of irrigable land and type of landscaping are not known, which have been shown previously to effect water consumption [38].

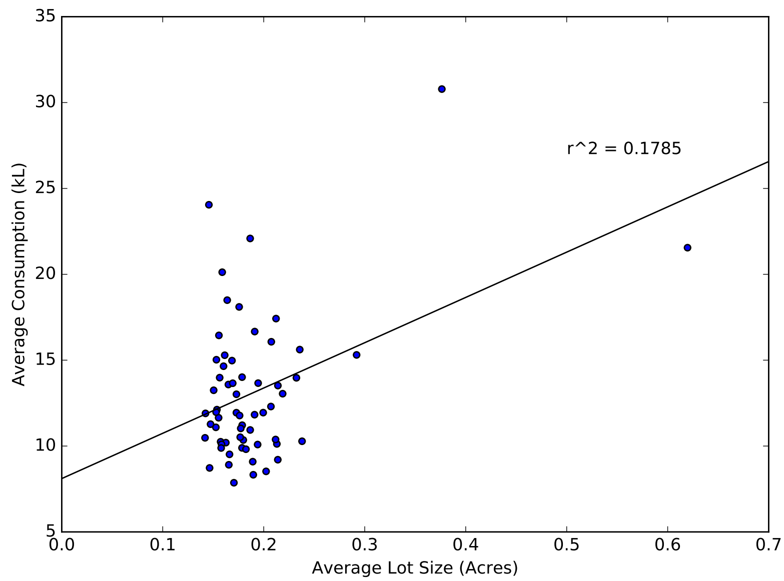

Figure 7.8: Correlation between lot size and outdoor water consumption by dissemination area

Figure 7.9 shows the correlation ($\rho = -0.0396, r = 0.145$) between the percentage of pools in a dissemination area and outdoor water consumption. The lack of correlation is consistent with previous work that shows pools contribute a relatively small amount to overall water consumption [3]. Additionally, only 139 (1.6%) of households in the dataset have pools. Water consumption for pool filling should not be a significant determinant of outdoor water consumption at the dissemination area level, given that most neighbourhoods have few pools, and even the largest percentage of pools in a dissemination area is only 9%.

While there are differences in per-household summer water consumption across dissemination areas (see Figure 4.1), the differences in outdoor water consumption are not strongly correlated with income, lot size, or rates of pool ownership as shown above. Additionally, the dissemination areas with the highest per-household water consumption contain few households, suggesting much of the variation at the dissemination area level is due to the variability in individual consumption levels.

7.5 Explanatory Factors at the Household Level

The major outdoor end-uses of water are irrigation and pool filling, and it is useful to know how characteristics affecting these end-uses vary by household.

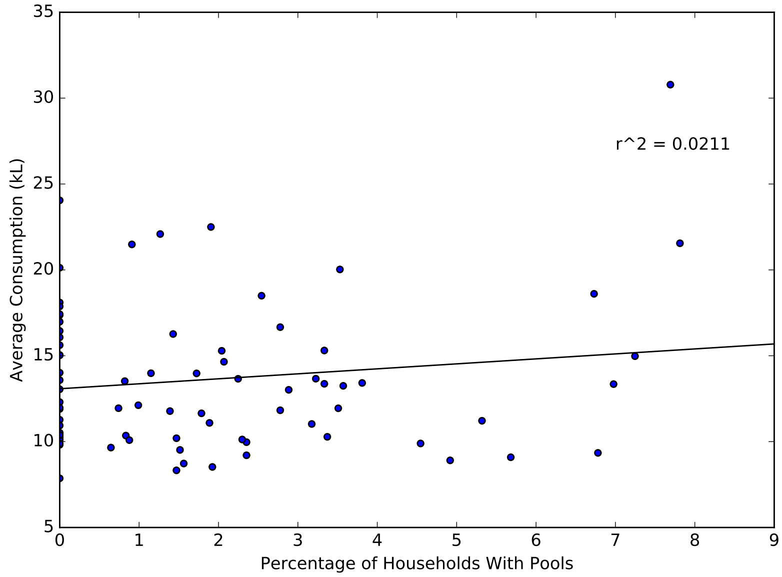

Figure 7.9: Correlation between pool ownership and outdoor water consumption by dissemination area

While rates of pool ownership do not significantly affect outdoor water consumption at the dissemination area level, there are significant differences in water consumption between individual households with and without pools. As shown in Figure 7.10, typical water consumption is much higher for households that have pools. The median consumption is 24.7 kL for households with pools and 4.7 kL for households without pools. Additionally, there is greater variability in water consumption for these households. These differences are obscured at the household level because of the relatively low rates of pool ownership in Abbotsford.

Assuming households used water for irrigation proportionally to lawn size, there should be some correlation between outdoor water consumption and lot size. However, there is essentially no correlation (at the household level) in the dataset. Because many households do not use significant amounts of outdoor water, the analysis is restricted to the 703 highest-consumption households which account for 50% of outdoor water consumption. Figure 7.11 shows almost no correlation ($\rho = 0.0817, r = 0.144$) between lot size and outdoor water consumption even for the households that use large amounts of outdoor water, where it might be expected that lot size would be a good measurement of irrigation needs. The lack of correlation may be because lot size is relatively consistent between households, with almost all households having relatively small lots. Additionally, no information about ground cover or type of landscaping is available, which has been shown to affect water consumption for irrigation [38].

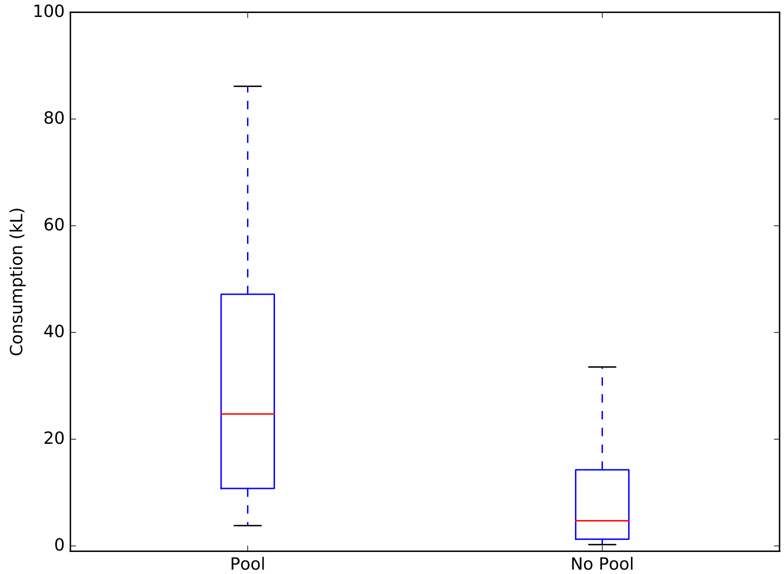

Figure 7.10: Variability of outdoor water consumptions for households with and without pools, showing the 5th, 25th, 50th, 75th, and 95th percentiles of consumption

7.6 Summary

In this section I showed some of the determinants of outdoor water consumption in Abbotsford, British Columbia. As in previous work, water consumption was shown to be sensitive to the occurrence of rainfall, and to vary somewhat based on temperature. Additionally, outdoor water consumption is extremely variable between households, but is not strongly related to demographic or property variables (besides the presence of a pool at the household level). This individual variability is an important consideration for demand management, and for the predictability of outdoor water consumption. In the next section I summarize the contributions of my thesis and discuss related work.

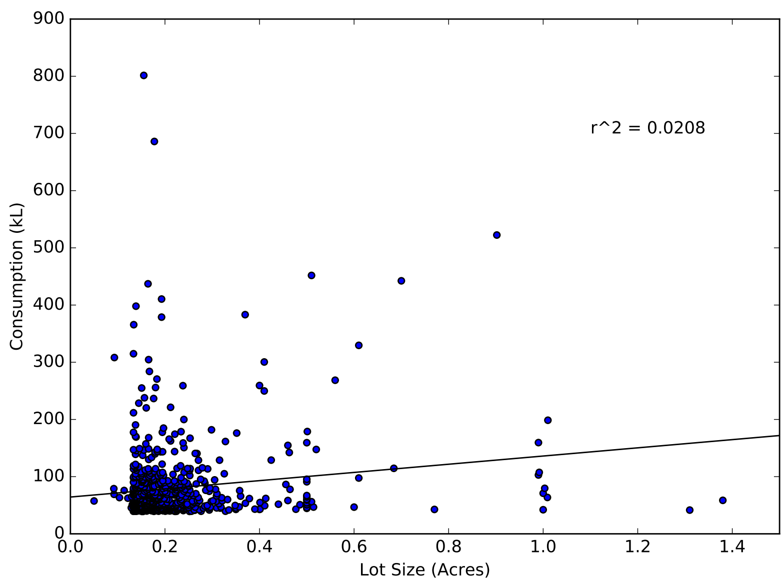

Figure 7.11: Correlation between lot size and outdoor water consumption by household, for households that use in the top 50 % of outdoor water consumption

Chapter 8

Conclusion and Future Work

This work has shown that outdoor water consumption can be identified using a simple upper threshold on probable indoor water consumption. Identifying outdoor consumption is important for demand management and reducing peak demand. Outdoor consumption estimated using this method contributes significantly to peak day demand, when compared to the contribution of outdoor use to average day consumption. I also showed that this threshold is relatively robust, and produces a reasonable estimate of outdoor consumption even if it is slightly modified, allowing for choosing a trade-off between identifying more outdoor water consumption and not miscategorizing indoor consumption. This method is useful because it uses data already collected by smart meters, and does not require additional equipment or higher-frequency recording.

The models developed for predicting outdoor consumption show some ability to forecast next-hour water consumption, but the accuracy leaves room for improvement. Short-term prediction of outdoor water consumption appears to be a difficult problem, because of its variability compared to total consumption. Additionally, outdoor water consumption varies significantly between households. This also contributes to the difficulty in predicting consumption for small neighbourhoods, because the average is significantly affected by individual behaviour.

Finally, the variability of outdoor water consumption between households and the lack of correlation between outdoor water consumption and demographic variables are important considerations for targeting conservation efforts. Because a small number of households consume very large amounts of outdoor water, significant reductions in demand could be achieved by targeting these households directly. In contrast, efforts targeted by demographic factors or by neighbourhood would likely be less useful. The ability to estimate outdoor consumption per household can therefore be an important tool for managing peak demand.

8.1 Future Work

While outdoor water consumption in Abbotsford is not affected by the demographic factors evaluated in this work, it is not clear if there are other factors which contribute to household-level variability in outdoor water consumption. Future work could focus on evaluating more variables such as the amount of irrigable area for each house, and factors such as attitudes to conservation.

The main limitations of this work involve the accuracy of the predictive models. While short-term predictions of outdoor water consumption appear to be a difficult problem, such predictions would be useful for day-to-day management and planning watering restrictions. Future work could involve more spatial aggregation to determine how large neighbourhoods must be to allow accurate prediction. In particular, daily prediction of water consumption would be more useful for planning watering restrictions, and this is an important problem for future research.

References

[1] D. Carboni, A. Gluhak, J. A. McCann, and T. H. Beach, "Contextualising water use in residential settings: A survey of non-intrusive techniques and approaches," *Sensors*, vol. 16, no. 5, p. 738, 2016.

[2] G. Cole and R. Stewart, "Smart meter enabled disaggregation of urban peak water demand: Precursor to effective urban water planning," *Urban Water Journal*, vol. 10, no. 3, pp. 174–194, 2013.

[3] M. Loh and P. Coghlan, *Domestic water use study in Perth, Western Australia, 1998-2001.* Water Corporation Perth, 2003.

[4] R. M. Willis, R. A. Stewart, K. Panuwatwanich, P. R. Williams, and A. L. Hollingsworth, "Quantifying the influence of environmental and water conservation attitudes on household end use water consumption," *Journal of Environmental Management*, vol. 92, no. 8, pp. 1996–2009, 2011.

[5] Environment Canada, "2011 municipal water use report : Municipal water use 2009 statistics," 2011.

[6] R. B. Billings and C. V. Jones, *Forecasting Urban Water Demand.* American Water Works Association, 2011.

[7] A. MacDonald, "Water needs: The demand for water," *Report of the Royal Society of Chemistry. University of Leeds*, 2007.

[8] P. W. Mayer, W. B. DeOreo, E. M. Opitz, J. C. Kiefer, W. Y. Davis, B. Dziegielewski, and J. O. Nelson, "Residential end uses of water," 1999.

[9] E. Donkor, T. Mazzuchi, R. Soyer, and J. Roberson, "Urban water demand forecasting: Review of methods and models," *Journal of Water Resources Planning and Management*, vol. 140, no. 2, pp. 146–159, 2012.

[10] L. House-Peters and H. Chang, "Urban water demand modeling: Review of concepts, methods, and organizing principles," *Water Resources Research*, vol. 47, no. 5, 2011.

[11] A. Wakeford, "2011 water conservation summary," Tech. Rep. WSC 1-2012, Abbotsford/Misson Water & Sewer Commission, 2012.

[12] K. Filiatrault, "2013 sprinkling restrictions," Tech. Rep. WSC 21-2013, Abbotsford/Misson Water & Sewer Commission, 2013.

[13] K. Filiatrault and T. Kyle, "Smart meter update," Tech. Rep. ENG 45-2011, Abbotsford/Misson Water & Sewer Commission, 2013.

[14] G. Hauber-Davidson and E. Idris, "Smart water metering," *Water*, vol. 33, no. 3, pp. 38–41, 2006.

[15] T. Boyle, D. Giurco, P. Mukheibir, A. Liu, C. Moy, S. White, and R. Stewart, "Intelligent metering for urban water: A review," *Water*, vol. 5, no. 3, pp. 1052–1081, 2013.

[16] R. A. Stewart, R. Willis, D. Giurco, K. Panuwatwanich, and G. Capati, "Web-based knowledge management system: linking smart metering to the future of urban water planning," *Australian Planner*, vol. 47, no. 2, pp. 66–74, 2010.

[17] T. Britton, G. Cole, R. Stewart, and D. Wiskar, "Remote diagnosis of leakage in residential households," *Journal of Australian Water Association*, vol. 35, no. 6, pp. 89–93, 2008.

[18] G. Bontempi, S. Taieb, and Y. Le Borgne, "Machine learning strategies for time series forecasting," in *Business Intelligence*, pp. 62–77, Springer, 2013.

[19] J. Friedman, T. Hastie, and R. Tibshirani, *The Elements of Statistical Learning*, vol. 1. Springer series in statistics New York, 2001.

[20] N. Japkowicz and M. Shah, *Evaluating Learning Algorithms: A Classification Perspective*. Cambridge University Press, 2011.

[21] L. Breiman, J. Friedman, C. J. Stone, and R. A. Olshen, *Classification and Regression Trees*. CRC press, 1984.

[22] J. H. Friedman, "Greedy function approximation: A gradient boosting machine," *Annals of statistics*, pp. 1189–1232, 2001.

[23] W. B. DeOreo, J. P. Heaney, and P. W. Mayer, "Flow trace analysis to assess water use," *American Water Works Association. Journal*, vol. 88, no. 1, p. 79, 1996.

[24] P. Mayer, W. DeOreo, and M. Alexander, "Trace wizard water uses analysis software," *Aquacraft*, 1999.

[25] "Identiflow for microcomponent analysis," *Water Research Centre*, 2014.

[26] J. Fogarty, C. Au, and S. E. Hudson, "Sensing from the basement: A feasibility study of unobtrusive and low-cost home activity recognition," in *Proceedings of the 19th Annual ACM Symposium on User Interface Software and Technology*, pp. 91–100, ACM, 2006.

[27] J. Froehlich, E. Larson, E. Saba, T. Campbell, L. Atlas, J. Fogarty, and S. Patel, "A longitudinal study of pressure sensing to infer real-world water usage events in the home," in *International Conference on Pervasive Computing*, pp. 50–69, Springer, 2011.

[28] E. Larson, J. Froehlich, T. Campbell, C. Haggerty, L. Atlas, J. Fogarty, and S. N. Patel, "Disaggregated water sensing from a single, pressure-based sensor: An extended analysis of hydrosense using staged experiments," *Pervasive and Mobile Computing*, vol. 8, no. 1, pp. 82–102, 2012.

[29] Y. Kim, T. Schmid, Z. M. Charbiwala, J. Friedman, and M. B. Srivastava, "Nawms: Nonintrusive autonomous water monitoring system," in *Proceedings of the 6th ACM Conference on Embedded Network Sensor Systems*, pp. 309–322, ACM, 2008.

[30] V. Srinivasan, J. Stankovic, and K. Whitehouse, "Watersense: Water flow disaggregation using motion sensors," in *Proceedings of the Third ACM Workshop on Embedded Sensing Systems for Energy-Efficiency in Buildings*, pp. 19–24, ACM, 2011.

[31] D. R. Maidment, S.-P. Miaou, and M. M. Crawford, "Transfer function models of daily urban water use," *Water Resources Research*, vol. 21, no. 4, pp. 425–432, 1985.

[32] S. L. Zhou, T. A. McMahon, A. Walton, and J. Lewis, "Forecasting daily urban water demand: A case study of Melbourne," *Journal of Hydrology*, vol. 236, no. 3-4, pp. 153–164, 2000.

[33] L. House-Peters, B. Pratt, and H. Chang, "Effects of urban spatial structure, sociodemographics, and climate on residential water consumption in Hillsboro, Oregon," *JAWRA Journal of the American Water Resources Association*, vol. 46, no. 3, pp. 461–472, 2010.

[34] G. J. Syme, Q. Shao, M. Po, and E. Campbell, "Predicting and understanding home garden water use," *Landscape and Urban Planning*, vol. 68, no. 1, pp. 121–128, 2004.

[35] A. A. Makki, R. A. Stewart, K. Panuwatwanich, and C. Beal, "Revealing the determinants of shower water end use consumption: Enabling better targeted urban water conservation strategies," *Journal of Cleaner Production*, vol. 60, pp. 129–146, 2013.

[36] S. Gato, N. Jayasuriya, and P. Roberts, "Temperature and rainfall thresholds for base use urban water demand modelling," *Journal of Hydrology*, vol. 337, no. 3, pp. 364–376, 2007.

[37] R. C. Balling, P. Gober, and N. Jones, "Sensitivity of residential water consumption to variations in climate: An intraurban analysis of Phoenix, Arizona," *Water Resources Research*, vol. 44, no. 10, 2008.

[38] B. Breyer, H. Chang, and G. Parandvash, "Land-use, temperature, and single-family residential water use patterns in Portland, Oregon and Phoenix, Arizona," *Applied Geography*, vol. 35, no. 1, pp. 142–151, 2012.

[39] A. Castledine, K. Moeltner, M. K. Price, and S. Stoddard, "Free to choose: Promoting conservation by relaxing outdoor watering restrictions," *Journal of Economic Behavior & Organization*, vol. 107, pp. 324–343, 2014.

[40] J. Bougadis, K. Adamowski, and R. Diduch, "Short-term municipal water demand forecasting," *Hydrological Processes*, vol. 19, no. 1, pp. 137–148, 2005.

[41] J. Adamowski and C. Karapataki, "Comparison of multivariate regression and artificial neural networks for peak urban water-demand forecasting: Evaluation of different ANN learning algorithms," *Journal of Hydrologic Engineering*, vol. 15, no. 10, pp. 729–743, 2010.

[42] B. Taylor, "Predicting normalised monthly patterns of domestic external water demand using rainfall and temperature data," *Water Science and Technology: Water Supply*, vol. 12, no. 2, pp. 168–178, 2012.

[43] D. R. Maidment and S.-P. Miaou, "Daily water use in nine cities," *Water Resources Research*, vol. 22, no. 6, pp. 845–851, 1986.

[44] M. Herrera, L. Torgo, J. Izquierdo, and R. Pérez-García, "Predictive models for forecasting hourly urban water demand," *Journal of Hydrology*, vol. 387, no. 1, pp. 141–150, 2010.

[45] A. Jain, A. K. Varshney, and U. C. Joshi, "Short-term water demand forecast modelling at IIT Kanpur using artificial neural networks," *Water Resources Management*, vol. 15, no. 5, pp. 299–321, 2001.

[46] D. Walker, E. Creaco, L. Vamvakeridou-Lyroudia, R. Farmani, Z. Kapelan, and D. Savić, "Forecasting domestic water consumption from smart meter readings using statistical methods and artificial neural networks," *Procedia Engineering*, vol. 119, pp. 1419–1428, 2015.

[47] M. Nasseri, A. Moeini, and M. Tabesh, "Forecasting monthly urban water demand using extended Kalman filter and genetic programming," *Expert Systems with Applications*, vol. 38, no. 6, pp. 7387–7395, 2011.

[48] P. Cutore, A. Campisano, Z. Kapelan, C. Modica, and D. Savic, "Probabilistic prediction of urban water consumption using the SCEM-UA algorithm," *Urban Water Journal*, vol. 5, no. 2, pp. 125–132, 2008.

[49] S. Zhou, T. McMahon, A. Walton, and J. Lewis, "Forecasting operational demand for an urban water supply zone," *Journal of Hydrology*, vol. 259, no. 1, pp. 189–202, 2002.

[50] J. Caiado, "Performance of combined double seasonal univariate time series models for forecasting water demand," *Journal of Hydrologic Engineering*, vol. 15, no. 3, pp. 215–222, 2009.

[51] D. R. Maidment and E. Parzen, "Cascade model of monthly municipal water use," *Water Resources Research*, vol. 20, no. 1, pp. 15–23, 1984.

[52] S. Praskievicz and H. Chang, "Identifying the relationships between urban water consumption and weather variables in Seoul, Korea," *Physical Geography*, vol. 30, no. 4, pp. 324–337, 2009.

[53] J. S. Wong, Q. Zhang, and Y. D. Chen, "Statistical modeling of daily urban water consumption in hong kong: Trend, changing patterns, and forecast," *Water Resources Research*, vol. 46, no. 3, 2010.

[54] S.-P. Miaou, "A class of time series urban water demand models with nonlinear climatic effects," *Water Resources Research*, vol. 26, no. 2, pp. 169–178, 1990.

[55] A. Akuoko-Asibey, L. Nkemdirim, and D. Draper, "The impacts of climatic variables on seasonal water consumption in calgary, alberta," *Canadian Water Resources Journal*, vol. 18, no. 2, pp. 107–116, 1993.

[56] R. C. Balling Jr and P. Gober, "Climate variability and residential water use in the city of phoenix, arizona," *Journal of Applied Meteorology and Climatology*, vol. 46, no. 7, pp. 1130–1137, 2007.

[57] H. M. Corbella and D. S. Pujol, "What lies behind domestic water use? a review essay on the drivers of domestic water consumption," *Boletín de la Asociación de Geógrafos Españoles*, vol. 50, pp. 297–314, 2009.

[58] S. L. Harlan, S. T. Yabiku, L. Larsen, and A. J. Brazel, "Household water consumption in an arid city: Affluence, affordance, and attitudes," *Society and Natural Resources*, vol. 22, no. 8, pp. 691–709, 2009.

[59] A. Polebitski and R. Palmer, "Seasonal residential water demand forecasting for census tracts," *Journal of Water Resources Planning and Management*, vol. 136, no. 1, pp. 27–36, 2009.

[60] H. Chang, S. Praskievicz, and H. Parandvash, "Sensitivity of urban water consumption to weather and climate variability at multiple temporal scales: The case of Portland, Oregon," *International Journal of Geospatial and Environmental Research*, vol. 1, no. 1, 2014.

[61] K. Kappel and T. Grechenig, "Show-me: Water consumption at a glance to promote water conservation in the shower," in *Proceedings of the 4th International Conference on Persuasive Technology*, p. 26, ACM, 2009.

[62] W. B. DeOreo, P. W. Mayer, B. Dziegielewski, and J. Kiefer, "Residential end uses of water, version 2," 2016.

[63] A. C. Worthington and M. Hoffman, “An empirical survey of residential water demand modelling,” *Journal of Economic Surveys*, vol. 22, no. 5, pp. 842–871, 2008.

[64] I. Pardoe, *Applied Regression Modelling.* New York: Wiley & Sons, 2012.

[65] E. A. Wentz and P. Gober, “Determinants of small-area water consumption for the city of Phoenix, Arizona,” *Water Resources Management*, vol. 21, no. 11, pp. 1849–1863, 2007.