Improved Artificial Neural Network Models for Predicting Hourly Water Consumption

by

Li (Steven) Wang

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 water meter devices are now widely installed in single family residences, allowing water consumption data to be collected at a high resolution from both the temporal and spatial perspectives. Such data allows improved prediction of future water consumption—an important task for water utilities as they manage the water supply. The dataset in this thesis consists of hourly water consumption data from the 9,045 single-family residences in Abbotsford, British Columbia from September 2012 to August 2013. This research focuses on predicting hourly water consumption by using improved artificial neural network (ANN) models and makes five main contributions. The first contribution is accurately predicting hourly water consumption at a finer spatial and temporal scale than previous work. The second contribution is gathering and studying a wide variety of datasets and related features for predicting future water consumption. In addition to water consumption data, daily weather information, demographic information, property information and date information during the same period of time are collected in the raw dataset. The third contribution is to systematically perform feature selection, an important step in building machine learning models but one that is absent from previous work on predicting water consumption. For different experiment criteria, customized feature sets assist the corresponding models to accurately predict the hourly usages. The fourth contribution is to improve prediction accuracy by building separate models for weekday and weekend prediction. Residents consume water in different patterns between weekdays and weekends. By tackling the predictions separately, better performance can be achieved with less complicated models. Lastly, this research investigates the performance of multi-hidden-layer ANN models versus single-hidden-layer models. Although, single-hidden-layer models are sufficient in theory, we show that multi-hidden-layer ANNs can lead to improved performance.

Acknowledgements

First, I would like to give special thanks to my advisor, Professor Peter van Beek, for his great advice, encouragement and dedication. Thanks to Valerie Platsko for her achievements on aggregation of neighborhood data based on demographic information. Last but not least, I would like to appreciate Professors Mei Nagappan and Professor Yaoliang Yu for their valuable suggestions and feedbacks after my thesis presentation.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

In this chapter, I informally introduce the problem addressed in this thesis: predicting short-term residential water consumption by using machine learning techniques. Moreover, the contributions of the thesis and the organization of the thesis are demonstrated.

1.1 Problem

Residential water consumption predictions are important for water utilities in both short-term and long-term water supply plans. Short-term supply plans ensure there is sufficient water to support residents' day-to-day consumption; meanwhile, the water utility company minimizes supply costs. In addition, another benefit of short-term prediction is detecting water leakage within a short period of time. For example, Britton et al. [16] proposed a strategy enabling rapid and effective post-meter leakage identification and managing water loss by using hourly smart metering and prediction data.

The term "finer grid", which is used throughout this thesis, is defined as high resolution data in both the spatial and temporal dimensions. For the spatial dimension, a finer grid refers to a higher resolution in a target group of people from the location perspective. For example, predictions of the water consumption of a city are spatially finer than predictions of the entire country and predictions in districts are in a finer grid when comparing with city level predictions. For the temporal dimension, a finer grid refers to a higher resolution in time. For example, daily consumption predictions are finer than the monthly predictions and hourly consumption predictions are finer than daily predictions.

Smart meter devices have been installed for all water clients in the city of Abbotsford, British Columbia, and water consumption is recorded on an hourly basis. Our initial dataset consists of over 20,000 clients' hourly water usage during a one year period from September 2012 to August 2013. Our focus here is on single-family residences, and after a proper data cleaning and preparation process, 9,045 residents' hourly consumption are targeted for experiments. Although, it proves infeasible to predict at the household level, in this thesis water usage is accurately predicted at the dissemination area (census tract) level, a finer grid than previous work, by using various artificial neural network (ANN) models.

1.2 Contributions

The first contribution of this thesis is predicting the water consumption at finer grids. Although the ideal scenario is predicting hourly water consumption at the household level, this thesis targeted hourly predictions at the dissemination area (census tract) level due to the high variances of single family consumption, which makes the prediction infeasible. Taking Saturdays 10:00 am data over the year as an example, in a dissemination area with 140 residents, the variance of the area hourly usage is 3.3 liters. However, 35 residents in this area have individual variance above 4.0 liters with the highest individual variance reaching approximately 40.0 liters. On the other hand, comparing with the predictions at the city and district levels, predicting at the dissemination area level raises the spatial resolution to a finer grid. While predicting a small population group's hourly usage, each individual instance plays an important role. Any abnormal usage at a certain time of a single family can result in significant model adjustments and notable impacts on the overall performance. Therefore, the most significant contribution of this work is predicting hourly water consumption at the dissemination area level with high accuracy.

The second contribution of this thesis is engaging various related datasets in the prediction. Urban families utilize water for different purposes and there are different factors impacting the amount of the consumption for each purpose. In general, water consumption is categorized into two broad subsets: indoor and outdoor consumption. Indoor consumption refers, for example, to usage in showering, drinking and cleaning; while, the outdoor consumption refers, for example, to usage in irrigation, pool filling and car washing. Both indoor and outdoor water consumption are impacted by many factors. Therefore, this thesis leverages as much information as possible to implement the predictive models in order to improve model performance. (All the features except the historical water consumption information are listed in Appendix A.) In contrast to previous work, this thesis

involves many datasets including previous water consumption data, weather information, date information, demographic information and property information. Hence, another contribution of this thesis is engaging large-scale related datasets into the model to increase the model performance.

The third contribution of this thesis is determining suitable feature sets for different experimental scenarios. As mentioned above, various datasets are engaged in this thesis and each dataset introduces a set of features. Consequently, a large number of features are available when building a model. Making decisions on selecting effective features becomes very challenging. Guyon and Elisseeff [26] suggest that feature selection is a process that improves prediction performance, provides faster and more cost-effective predictors and provides better data and model understanding. In contrast to previous work, this thesis first builds a model with features based on the previous work and considers it as a baseline. Then models with feature selection are implemented and their performance is compared with the baseline model. Therefore, the third contribution of this thesis is providing detailed feature selection for models and showing that models with feature selection outperform ones with no feature selection.

The fourth contribution of this thesis is splitting the predictive model into two models—one for weekdays and one for weekends—to improve the overall prediction accuracy. Weekday and weekend water consumption follow different patterns. Not only the peak hours are different between weekdays and weekends, but also the amount of peak hour consumption are remarkably disparate from each other. This leads to the question of whether separate models for weekdays and weekends can significantly improve the model performance. When comparing the models' overall performance, the ones with weekdays and weekends separation outperform the others. Hence, the model separation is considered as one of the significant contributions in this thesis.

The final contribution of this thesis is engaging multi-layer neural network models; i.e., ANNs with more than one hidden layer. Although there is much evidence that deeper neural networks can outperform shallow networks, it is currently unknown whether multi-layer ANN models would outperform single-layer models in the context of water prediction. This thesis leverages the one day in advance hourly water experiments to compare models' performance on models with one, two and three hidden layers. Therefore, the last contribution of the thesis is identifying the best single or multi-hidden layer model(s).

1.3 Organization of the Thesis

In Chapter 2, I discuss previous work and achievements related to water prediction and artificial neural networks. Moreover, the achievements that this thesis is based on are emphasized such as which datasets are engaged, why splitting weekday and weekend models is useful, and why it is important to analyze the models' performance at peak hours. In Chapter 3, I present the data preparation procedures that are the foundation of this thesis. The data preparation includes data collection, data cleaning and data aggregation. In Chapter 4, the detailed feature selection process and results for all scenarios are presented, including baseline model; one hour ahead weekday and weekend models with and without predefined features; and one day ahead weekday and weekend models with single, two and three hidden layers. A best feature set is selected for each experiment set to optimize each model and make model implementations efficient. In Chapter 5, models based on the features from Chapter 4 are implemented. Moreover, one best model is proposed to represent each of the scenarios. In Chapter 6, I compare the alternative solutions determined from Chapter 5 from both model performance and structure perspectives. Moreover, the results are analyzed from peak seasons, peak hours, and weekdays and weekends aspects. In Chapter 7, the achievements of this thesis are summarized and potential future work that may improve the model prediction is suggested.

Chapter 2

Related Work

In this chapter, I review relevant previous work. First, I provide a brief introduction to artificial neural network (ANN) models, discuss previous work on applying ANN models to predicting water consumption and illustrate differences between single and multi-hidden layers ANN models. Second, I present previous work on forecasting water consumption in different temporal and spatial dimensions. Next, I provide an overview of previous research leveraging various datasets including weather, demographic and property information to predict water consumption. Then indoor and outdoor water consumption analysis are demonstrated. Last, I show two characteristics—peak hour usage and weekday/weekend separations—of water consumption, which are widely used to analyze residential water consumption patterns.

2.1 Artificial Neural Networks

In this section, I provide an overview of artificial neural network models which are referred to as ANNs for the rest of this thesis, demonstrate previous studies on ANNs predicting urban water consumption, and lastly present related work on multi-hidden layer ANNs.

This thesis uses ANNs to predict hourly water consumption. Hence, I first review ANNs' structures. There are three types of layers in ANNs: input layer, hidden layer and output layer. Each layer contains nodes designed to simulate neurons in a brain. Except for the output layer, nodes in each layer are fully connected to nodes in the next subsequent layer by directed edges. This means that every node in the current layer connects and only connects to all the nodes in the adjacent subsequent layer. Moreover, there is a weight assigned to each directed edge in order to calculate output results.

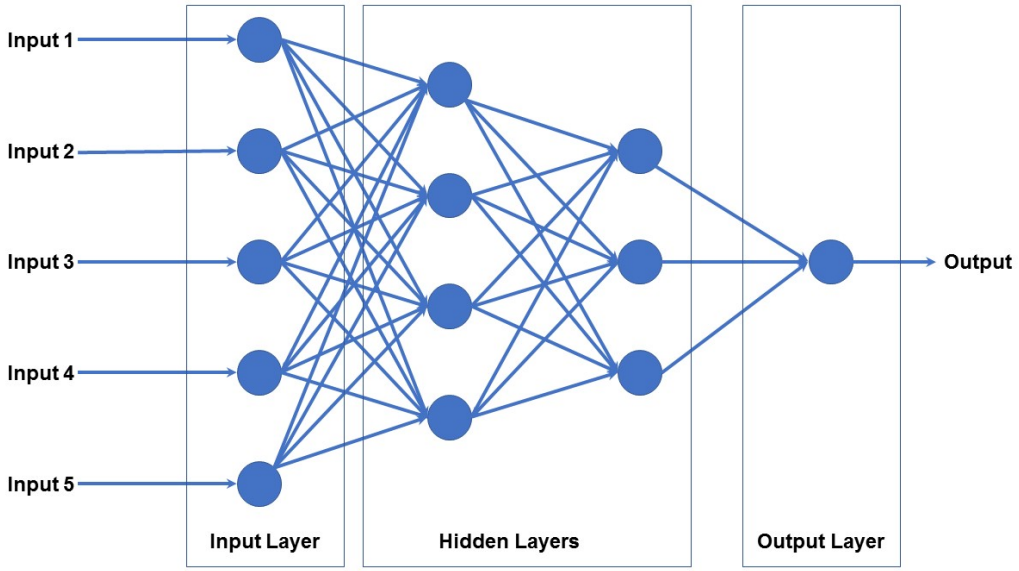


Figure 2.1: Artificial neural network model structure.

Figure 2.1 demonstrates a "deep" neural network model with two hidden layers. A first type of layer is an input layer. Nodes in the input layer represent the input signals received by the ANN. In this thesis, they refer to values of the corresponding features selected from historical water consumption, daily weather, demographic, date and property information datasets. The second type of layer is a hidden layer. ANNs can maintain zero, one or multiple hidden layers. Kim [37] divides ANNs into three categories based on the number of hidden layers: single layer ANNs, which retain no hidden layers, shallow ANNs, which retain only one hidden layer and deep ANNs, which retain more than one hidden layer. In this thesis, the last two types of models are implemented and they are referred as single hidden layer ANNs and multi-hidden layer ANNs. In addition to nodes and edges in hidden layers, there is one more component called the activation function. For each node in a hidden layer, the node first calculates a weighted sum defined as,

$$y = wx + b,$$

where w is a vector that represents weights of all directed edges incoming to a node, x is a vector of the corresponding input values that are fed into the node, and b is a bias term. The weighted sum is passed to an activation function and an output value is calculated. There are different options for activation functions. The sigmoid function is selected in

this thesis,

$$\sigma(x) = \frac{1}{1 + e^{-x}}.$$

Finally, the value calculated by the activation function is utilized as the inputs of the immediate subsequent layer. The last type of layer is an output layer. Although Figure 2.1 contains only one output node, there could be multiple nodes in this layer. This thesis focuses on predicting hourly water consumption; therefore, only one output node is required. The output calculation in this layer is the same as the one defined in hidden layers. As the output value is not restricted from 0 to 1, the activation function is

$$f(x) = x,$$

which means that the weighted sum is the output.

2.1.1 ANNs in Water Usage Prediction

ANNs have been well investigated in recent decades for water consumption prediction. Firat et al. [21] investigate different types of ANNs for monthly urban water consumption of the city of Izmir based on historical monthly water consumption data. Finer grid data has also been used in ANNs for high resolution water consumption forecasts. Walker et al. [53] evaluate different ANNs on domestic water consumption. The study takes nine residents of Greece with a high resolution of data and predicts water consumption on an hourly basis. Comparing with other models, ANNs demonstrate performance advantages over others, especially in water usage prediction. Adamowski et al. [1] implement city wide daily water demand prediction models for the city of Montreal by using nonlinear regression, auto-regressive integrated moving average, artificial neural network and wavelet artificial neural network methods. Bougadis et al. [15] compare the performance of ANNs, regression models and time series models for water consumption of the city of Ottawa on a weekly basis. Both suggest that ANNs dominate the performance of other machine learning methods. Moreover, Gagliardi et al. [23] compare a pattern based model with ANNs for short-term water demand forecasts. The results indicate that for small population groups, ANNs outperform other methods.

2.1.2 Single Hidden Layer Models and Multi-Hidden Layers

There has been much previous work investigating the performance of different numbers of hidden layers in ANNs. Although, Hecht-Nielsen [28] indicates that three layer models

should be able to describe all problems in theory, the optimal number of layers is an empirical question. Kumar et al. [38] investigate daily grass reference crop evapotranspiration and compare performance of ANNs with conventional methods. During the comparison, they discover that two hidden layer models attain lower performance than single hidden layer models for both training and validation sets when the number of learning cycles are the same. Although Panchal et al. [46] state that one hidden layer should be sufficient for almost all problems, experiments on employee retention data reveal that as soon as the number of hidden layers increases, the experimental results improve. Furthermore, Lee et al. [39] study applying ANN models to short-term load forecasts for a large power system. They encounter a situation where one hidden layer model cannot achieve their predetermined tolerance; however, a model with two hidden layers can. In this thesis, I study the performance of different numbers of hidden layers on water usage prediction.

2.2 Water Usage Prediction at Different Spatial and Temporal Levels

As mentioned in Chapter 1, the finer grid is defined from both the spatial and temporal perspectives. From the spatial perspective, predictions can be from a country level to a single-family level; while from the temporal perspective, predictions can be from a monthly level to an hourly level. There has been much previous work on residential water prediction at different spatial and temporal levels. A comparison of water prediction in this research and previous work is shown in Table 2.1.

Cominola et al. [19] indicate that traditional research is based on low-resolution data to model water demands at the city or block scale by using time resolutions at or above daily levels. Ghiassi et al. [25] implement dynamic ANNs for monthly, weekly, daily and hourly water consumption predictions of San Jose city at high accuracy rates. However, daily, weekly and monthly models are at the city level and hourly water predictions are at the zone level, where a zone contains approximately 250,000 people. Thus, all the experiments are conducted at low resolutions from the spatial perspective. Jain and Varshney [33] investigate the performance of ANN, regression analysis and time series analysis models for the weekly water demand in Kanpur, India. Although the study area is in a relatively finer grid, the predictions are at the weekly level, which is considered as medium or low resolution from a temporal perspective. In contrast, Zhou et al. [58] forecast operational demands for an urban water supply zone which is a residential area and has a population of 35,000 persons, at the daily and hourly level. Thus, some of the experiments are conducted at a finer temporal resolution but at a low spatial resolution.

Table 2.1: Comparison of previous work versus this research on temporal and spatial dimensions.

	City	Block	Certain group	Dissemination
Minute		Liu et al. [41]		
Hourly			Ghiassi et al. [25] Zhou et al. [58] Beal and Stewart [12] Bennett et al. [13]	This research: one model to predict 111 dissemination areas
Daily	Cominola et al. [19] Ghiassi et al. [25]	Cominola et al. [19]		
Weekly	Cominola et al. [19] Jain and Varshney[33] Ghiassi et al. [25]	Cominola et al. [19]		
Monthly	Cominola et al. [19] Ghiassi et al. [25]	Cominola et al. [19]		

As smart meter devices are now widely installed, high-resolution data has become available. Researchers have leveraged those data to predict water consumption patterns at finer grids from both the spatial and temporal perspectives; however, the experimental datasets are relatively small. Liu et al. [41] investigate water end-usage of 141 residents located at Tea Garden suburbs, Australia by utilizing smart meter to collect water consumption data at 1-minute intervals. Beal and Stewart [12] study water end-usage that drives peak day demand and the associated hourly demand patterns by investigating 230 residences' smart meter data over 18 months. Bennett et al. [13] leverage daily end-use consumption information of 205 households living in South East Queensland region associated with the water stock efficiency and demographic data to implement ANNs predicting water consumption. The end use information includes toilets, clothes washers, showers, dishwashers, taps and total internal demands. In contrast to previous finer grid predictions, this research leverages only one model to predict 111 dissemination areas, each contains 30 to 178 single family residences, hourly water consumption and the model retains high accuracy rates.

2.3 Weather Information

Climatic variables—in particular, precipitation—are essential factors for predicting water consumption. Some research considers the amount of rainfall information more important than the number of occurrences; whereas, others consider occurrences more important. For example, Balling et al. [35] investigate the relationship between monthly water consumption and climatic variables by using data between 1995 and 2004 in Phoenix at the city level and conclude that water consumption generally increases when temperatures are above normal, precipitation is below normal and areas are in periods of drought. On the other hand, Martínez-Espińeira [44] explains domestic water consumption by using price, billing, climatic and sociodemographic variables. The research considers rainy days—number of precipitation days in a month—over the total number of precipitation events in a month. One achievement of this thesis is that I provide solid reasons on when one feature dominates the other based on effective feature selection algorithms. Both precipitation amount and occurrences are included in the initial feature set; however, whether it should be engaged as an input depends on the feature selection results for different criteria.

This thesis performs experiments with two goals: predicting one hour ahead and predicting one day ahead hourly. As weather information is considered highly relevant, daily weather information is also included in one day ahead predictions. That means future weather information is required. Fortunately, there is much previous work that provides evidence that one day ahead weather information can be predicted accurately. Mohandes et al. [45] compares autoregressive and ANN models by predicting daily wind speed, and shows that ANN models outperform autoregressive models and achieve high accuracy. For daily average temperature, Tasadduq et al. [52] utilize ANN models to predict hourly mean values of ambient temperature for 24 hours ahead. They select one-year continuous data for model training and validate model performance by using another three years' predictions, and again show that ANN models achieve high accuracy. Besides predicting each individual meteorological variable, Maqsood et al. [43] conduct a weather forecast in southern Saskatchewan for a 24 hours ahead scenario. The climatic variables temperature, wind speed and humidity are predicted in four different seasons. The results demonstrate that the correlation coefficients between the actual and predicted values are above 99%. With these achievements from previous research, this thesis leverages daily weather information even for 24 hours in advance predictions.

2.4 Demographic Information

When investigating water consumption, demographic information is also widely considered. The top three crucial factors in this feature set are householder ages, incomes and education levels. Clark and Finley [18] study the determinants of water conservation intention for 728 residents in Blagoevgard, Bulgaria. The study concludes that ages, incomes, resident types and presence of gardens significantly related to residents' intention to implement water conservation measures. While studying determinants of residential water demand in Germany, Schleich and Hillenbrand [50] find that every 1% annual growth rate in per capita income results in 6.5 liters more water consumption per day per capita. Furthermore, they observe that higher ages are associated with higher water usage. While investigating 132 residences water end-use behaviors in the Gold Coast, Australia, Willis et al. [54] categorize household income into lower middle, middle and upper middle classes. They conclude that upper middle class residents have less water conservation concerns.

2.5 Property Information

In addition to demographic information, resident property information is considered pivotal as well. Fox et al. [22] propose methodologies to classify household property information for water consumption forecasts. The features of properties considered are number of bedrooms, building types (for example, detached, semi-detached, and townhouse) and whether a property has a garden attached. Their research indicates that the more bedrooms, the more consumption is required, and garden presence significantly correlates with increased water usage. House-Peters et al. [30] use regression models to discover effective factors that impact householders water consumption at the census block level in Hillsboro, Oregon. The Hillsboro water district consists of 37 adjacent census blocks. Outdoor usage is found to be strongly related to education levels and house sizes. Specifically, clients whose usage is sensitive to climate changes are the well-educated residents living in expensive and large houses. Similarly, Chang et al. [17] study single family resident water consumption at the block level which is defined by U.S. Bureau of the Census in 2007, for the city of Portland, Oregon. The study concludes that building sizes and ages are the most important property variables explaining water consumption.

2.6 Indoor and Outdoor Water Consumption

In general, water usage can be categorized into indoor and outdoor consumption. Indoor consumption consists of day-to-day usage including showering, drinking, washing and cleaning. Outdoor consumption includes irrigation, gardening and swimming pool filling. Balling et al. [34] analyze the annual water consumption data in the city of Phoenix to understand associations between climate variabilities and residential water usage. They indicate that climate variables are pivotal factors for outdoor water consumption. Polebitski and Palmer [48] investigate single family bimonthly water consumption patterns for residents in Seattle and find that outdoor water consumption is primarily driven by the local weather, whereas indoor usage is impacted by residence sizes and densities of residents living in a property. One more distinct impact variable between indoor and outdoor usage is family income. Higher income families tend to consume more outdoor water during summer time; whereas, indoor usage is insensitive to income distinctions. Some previous work has considered indoor and outdoor usage separately. For example, Kenney et al. [36] work on residents' water consumption patterns in Aurora, Colorado. They differentiate indoor and outdoor usage and treat them as separate variables in their regression model.

2.7 Characteristics of Water Consumption

In water consumption studies, researchers normally divide usage analysis into different categories. Two of the most popular categorizations are to divide consumption into peak and off-peak hours and to separate weekday and weekend usage.

Peak hour water consumption is widely considered in previous work while analyzing water usage behaviors. Arbués et al. [9] in their work on estimating residential water demand indicate that peak hour consumption is less sensitive to price changes than off-peak hour usage. Moreover, in order to design a water supply system for the Western United States, Hughes [32] takes peak seasons, peak days and peak hours information into account so that the system ensures reservoir storage, pump plant sizes and the pipeline sizes sized for peak demand. Gargano et al. [24] demonstrate probabilistic models for peak hour usage and indicate that peak hour consumption predictions are pivotal for water supply system designs and managements.

Due to different water consumption patterns during weekdays and weekends, much previous work prefers to analyze weekday and weekend consumption behaviors separately. When Alvisi et al. [8] implement short-term pattern based models for hourly water demand

predictions, they observe variable diurnal patterns for weekdays and weekends. Similarly, Blokker et al.[14] consider weekdays and weekends separately while simulating residential end-use water demand. Moreover, they observe different relationships between water consumption and clients' ages and occupations. Weekday consumption retain strong associations with these variables; whereas, weekend consumption do not. In addition, Eslamian et al. [20] leverage a novel multi-regression model to forecast daily urban water consumption and demonstrate that water consumption increases significantly due to weekend factors. Moreover, Alpaydin [6] indicates that when a dataset has multiple classes following different distributions, choosing multiple model is considered as a better option comparing to a single model. Therefore, the weekday and weekend data is separated in this research and the predictions of each scenario are conducted independently.

2.8 Summary

In this chapter, I summarize relevant previous work. I first introduce ANNs and their achievements. From a data resolution perspectives, I present previous work on water usage prediction at different temporal and spatial dimensions. In contrast to previous works, this thesis predicts water consumption at the hourly level and at the dissemination area (census tract) level by engaging large scale single-family consumption data for the city of Abbotsford, British Columbia. Moreover, I discuss a variety of datasets for water usage forecasts in previous work. The datasets include weather, demographic and property information, which are determined as feature sets for this thesis. To understand how variables impact water consumption, I present previous work on indoor and outdoor water consumption analysis. Last, the analysis of peak hour usage and weekend and weekday separation in previous work is illustrated. This research imports the weekend and weekday separation strategy and analyzes models' performance in the peak-hour manner.

In the next chapter, I present how water consumption, weather, demographic, property and date information is collected, filtered and cleaned. Data is the foundation of this research as it is utilized by feature selection, and used in learning the predictive models and in performance analysis. Hence, ensuring data integrity and data cleansing is crucial.

Chapter 3

Data Collection and Data Cleaning

There are five datasets created in order to retain weather, water usage, timeline, demographic and assessment information for this research. Datasets are selected based on previous work. The target variable for this thesis is next hour water usage. Predictor variables are distributed in selected datasets such as daily temperature in the weather information, previous hour water usage information in water usage dataset, and public holiday information in timeline information dataset. Understanding the datasets will enable an understanding of water consumption patterns. Hence, each dataset will be presented in turn, prior to addressing model construction and selection.

3.1 Smart Water Meter Data

Hourly water usage information is time series data. From a time series analysis perspective, previous water usage should have a close relationship with current and future water demand. Herrera et al. [29] demonstrate the importance of previous hour, previous two hours and previous-week-same-hour water usage information to current hour water prediction models. Babel and Shinde [10] indicate that historical water demand information is one of the critical features for short term and long term water usage prediction models.

Smart-meter devices have been installed for each water customer in the city of Abbotsford, British Columbia. The meters record hourly water consumption and send data back to the utility company's servers. In this thesis, I focus on single-family residences and the main dataset used consists of hourly water consumption measurements for each household in Abbotsford, recorded from September 2012 to August 2013. Before using the dataset in

a machine learning context, it is important to ensure that the data is clean. Thus, I first filtered unnecessary data, cleaned dirty data and enriched missing data.

3.1.1 Data Cleaning

Data cleaning is an inevitable procedure that needs to be followed for data related projects. Zhang et al. [57] mention that due to impure data in the real world, data preparation is pivotal for building a high performance data mining system. Fox et al. [22] demonstrate a data cleaning process for water demand prediction. However, in their work, the cleaning process results in a small experimental dataset: only 566 out of 1555 sample instances are left after cleansing. In the following, I provide detailed information on data preparation of the single-family hourly water consumption dataset. As suggested by Zhang et al. [57], the data cleaning process involved removing outliers (un-repairable data), resolving data conflicts and imputation of missing data.

Step 1: Remove outliers.

Information for over 20,000 customers was provided initially, including 9,918 single-family residences. However, due to network and hardware issues, some customers' hourly consumption was not recorded correctly. For example, during peak hours (e.g., 6:00–8:00am on weekdays), the hourly consumption was recorded as zero whereas the next hour usage was recorded as exactly one cubic meter. The explanation for this exception is that accumulated water consumption information was transferred rather than hourly water usage information. There were 873 clients following this pattern and their water usage information was filtered out. After this data filtering process, the remaining 9,045 householders' hourly water usage information was saved into a repository.

Step 2: Resolve data conflicts.

The first conflict in the hourly water consumption data is that the recorded system time zone and the actual local time zone were discovered to be different. Adamowski [2] studies peak daily water usage in Ottawa and observes that residential demand increases significantly after 4:00pm, reaches a peak just before 9:00pm, and then gradually decreases. However, a plot of hourly water usage for the Abbotsford data indicated that the demand increased significantly after 9:00am, reached a peak around 2:00pm for weekdays, and then gradually decreased. This represented a shift of approximately seven hours from

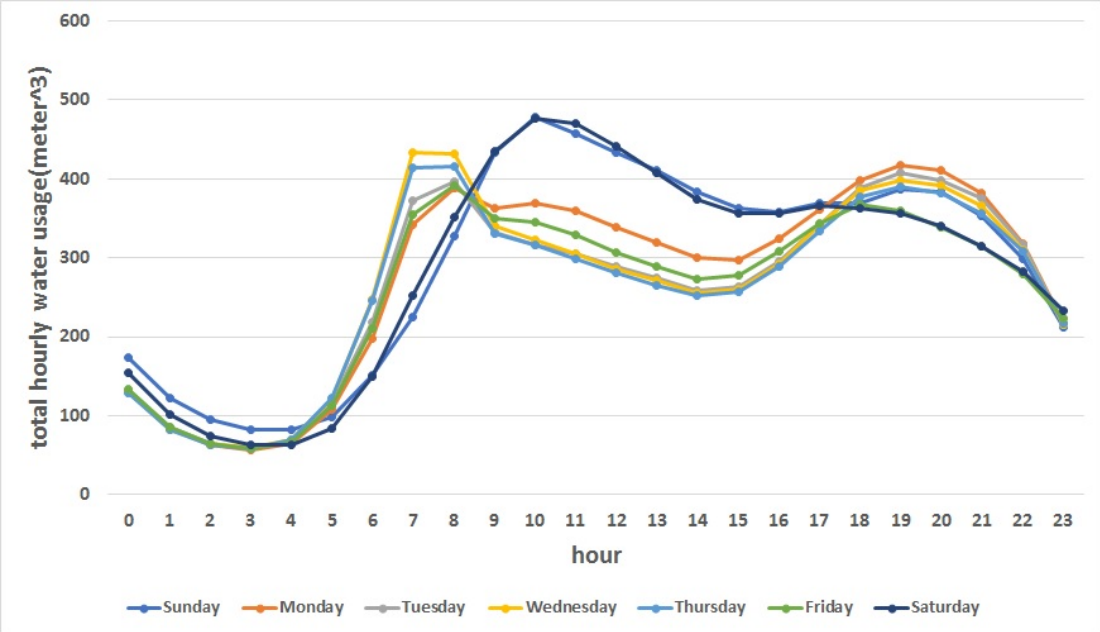


Figure 3.1: Hourly water usage for each day of the week after correcting for time zone.

previous work. Subsequently, it was discovered—and confirmed with the water utility in Abbotsford—that the hourly water consumption was recorded using the universal time zone, whereas Abbotsford is in the Pacific time zone. The solution to the conflict was to adjust the recorded times to the Pacific time zone. Thereafter all the peak hours match to the expected periods (see Figure 3.1). For example, during weekdays the peak hour in the evening is around 9pm.

The second conflict in the hourly water consumption data was due to daylight savings time. For our dataset, there were two times where the clock was adjusted due to daylight savings time: November 4, 2012 at 2:00am and March 10, 2013 at 1:00am. On November 4 at 2:00am the clock is turned backward to 1:00am with the result that there were two records for 1:00am. In contrast, there were no records for March 10 1:00am. For the duplicated records, mean value of the two hours is taken as the consumption. For the missing hour, it is treated as a missing hour (see below for how the missing data was imputed).

Step 3: Impute missing data.

After conducting the first two steps above, there are 8,664 hourly water usage data collected for each householder. Comparing to 8,760 hours, which are 24 hours per day and 365 days during the entire period, there are 96 missing records for each customer. Missing data is a consequence of hardware device maintenance. As indicated by the utility company, there are 4-time periods when water usage data is not collected due to the company's maintenance schedules. These time periods are listed in Table 3.1.

Table 3.1: Missing hours in water time series data for all single-family residences.

From	To	Hours
2013 Feb. 16 17:00	2013 Feb. 17 16:00	24 hours
2013 Mar. 9 17:00	2013 Mar. 10 17:00	25 hours including one hour switch
2013 Mar. 30 18:00	2013 Mar. 31 17:00	24 hours
2013 Jul. 27 18:00	2013 Jul. 28 17:00	24 hours

In order to address this data issue, estimated values are used to enrich the data. There are two different mechanisms applied and the optimal solution is selected at the end.

There are 97 hours with missing values listed in Table 3.1, 96 regular hours and one missing hour (March 10, 1:00 am) due to daylight saving time. This research evaluates two different approaches to estimate values and selects a best solution. Before presenting data enrichment, it is important to understand why the missing data cannot be ignored. Historical hourly consumption is an important factor for future water usage predictions. For example, Bakker and Duist [11] engage previous day water consumption as parameters to their adaptive heuristic model to predict one-day ahead water usage. Herrera and Torgob [29] leverage previous week same hour water usage information to predict next hour water usage in multiple models. Based on previous work, this thesis engages the previous entire weeks' hourly consumption as a raw feature set and takes them as inputs of feature selection to ensure all possible effective predictor variables are considered. In consecutive hours, even though there is only one-hour usage missing, the consequence is that the entire data for the following weeks is lost in the experiment dataset. Enriching missing values benefits the feature selection and model implementation sections, and thus retains data integrity.

Different clients may follow different water consumption patterns. Although this research aggregates water usage at the dissemination area level, significant differences be-

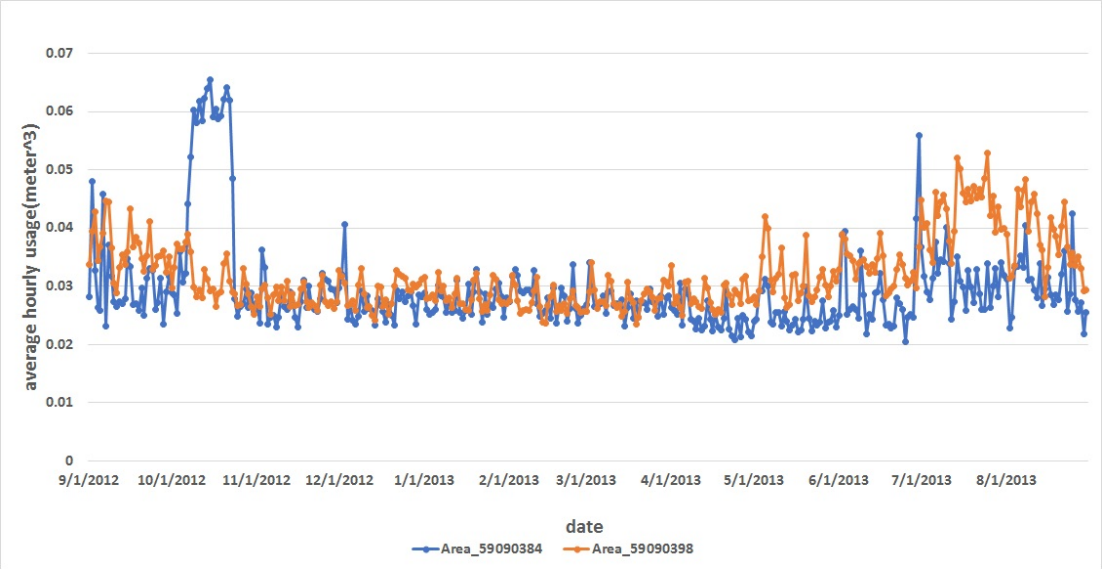


Figure 3.2: Comparison of average hourly water usage for two dissemination areas.

tween dissemination areas persist. Figure 3.2 shows the average hourly usage for two dissemination areas, where the population is aggregated at the dissemination area level. The two consumption patterns are remarkably different from each other, especially during summer periods. Han and Kamber [27] introduce several efficient ways to address missing data issues. Two approaches are selected for gap filling in the present research: mean value and regression tree prediction. Our first approach is the mean value process. As Peters and Chang [31] and Bakker and Duist [11] demonstrate, weekly water consumption follow certain patterns, and so do hourly water usage for a particular hour of a day. By following this conclusion, for each householder, water usage is divided into 168 subgroups representing 7 days per week and 24 hours per day. All usage data is assigned to one of the groups, including missing data. Thereafter, the mean value of each group is calculated excluding missing data. At the end, all missing hourly water usage are filled by the mean value of their corresponding group. Our second approach follows a regression tree strategy. Witten and Frank [55] define a regression tree as a decision tree with averaged numeric values at the leaves. Lewis [40] demonstrates that a regression tree is an efficient procedure for low dimensional data predictions. To fill the gaps in current hour usage, previous one-hour usage, previous two-hours usage, and previous week same hour usage are selected as input features. Thereafter, the problem is transformed to using three parameters to predict next hour water usage.

Experiments are performed on a small dataset consisting of 168 randomly selected householders. As the mean values approach is dominated by the regression tree approach, the latter was chosen for data imputation. Once the regression tree model is constructed, imputation takes place and values are filled by following hour indices in an ascending order as latter missing hours' gap filling may need former gap filled values.

3.2 Data Grouping

This research focuses on predicting water consumption in a finer grid. Statistics Canada divides the city of Abbotsford into 158 adjacent dissemination areas geographically. This research imports achievements from Platsko [47] by integrating single family address information with dissemination area information. Each address is assigned to only one dissemination area based on its longitude and latitude values.

The number of single family residents in each dissemination area varies from 1 to 178. Small sample sizes introduce challenges for prediction. Water consumption patterns for a small population are more unstable than for a larger population. For example, within a small group of residents, an outstanding hourly water usage for a single consumer has great impact on the group water usage. If a family is away on vacation, the usage at peak hours of a small group is pulled downwards significantly. This may introduce large prediction errors. Therefore, for those areas with a small number of single family residents (SFRs), this research amalgamates these residents with those of adjacent dissemination areas.

Since some of the dissemination areas have multiple neighbors, a method for determining which adjacent areas should be merged is needed. For this purpose, the 2011 National Household Survey data for Abbotsford and Mission from Statistics Canada is engaged. The dataset includes demographic and property information for each dissemination area. Moreover, the most likely dissemination areas which are determined by feature values in the dataset are merged together as needed.

Initially, there are more than 300 features for each dissemination area such as income tax, ages and education levels. In order to efficiently leverage these features, demographic information and building information are selected because they have been proven to be effective factors for water usage prediction. Aitken et al. [4] investigate relationships between weekly water consumption and a range of statistic variables such as property value and concludes that regression models consisting of clothes washing-machine loads per week, number of people per household and property value explain 60% of the variability of residential consumption. Liu and Savenijea [42] leverage ANNs with input variables consisting

of water price, house income and house size to predict water consumption per capita per day. An interesting conclusion is that around 60% of high-income single families prefer dinners at restaurants instead of at home, which leads to lower water consumption. Shandas and Rao [51] combine property sizes and ages with weather and day of week information to predict daily water consumption. They demonstrate that building sizes, ages and assessed values together explain about 20.7% of water consumption. Based on previous achievements, there are 13 features selected from three different categories: house structures, education levels and financial features.

Platsko [47] indicates neighborhoods of all dissemination areas in Abbotsford. Moreover, this thesis uses the k*th* nearest neighbors method to find the most similar neighbor to merge based on features listed above. Specifically, if an area contains less than a certain number of single family residences, it is merged to an adjacent region that retains the most similar feature information. MATLAB's `knnsearch` function is used. Furthermore, the Minkowski method is used as the search method in this process. In order to optimize the solution, all the prerequisite variables values are normalized.

The threshold for the minimum number of single family residences in the area is set at 30. If a region contains less than 30 SFR, then it will be merged to the most similar adjacent area. The merged region's feature values will be the combined value of each individual area. The procedure repeats until all regions contain more than 30 SFRs. When two areas are merged, the neighborhoods' information as well as the statistical information is combined.

There are 52 regions containing less than 30 SFRs. It takes only one round of merging to ensure that all areas retain more than the threshold population. Lastly, every single family resident is assigned to a dissemination area and hourly water usage are aggregated for each resulting area.

3.3 Weather Data

Resident water consumption have shown a close relationship with weather information (see Chapter 2). The website http://abbotsfordwx.com/ provides detailed daily weather information between September 1, 2012 and August 30, 2013 with no missing information. Average daily temperature, barometer, windspeed and accumulated rainfall so far this month are directly captured by the source. Daily rainfall information is reassembled by using the current accumulated amount minus the previous day's value except the first day of a month. Over all weather features, temperature information keeps the closest relationship with daily consumption values. The correlation coefficient between temperature and water

consumption is 0.77. Figure 3.3 suggests that daily water consumption increases as daily temperature increases. Besides these variables, the rainfall occurrence has been considered as a crucial factor (see Chapter 2). In this experiment set, the rainfall occurrence has strong positive associations with water consumption, especially in the summer (from May 1 to August 30). Over the entire year, the correlation coefficient is 0.43; however, the number rises to 0.60 during the summer because of massive consumption for irrigation and gardening. Compared to the rainfall occurrence, the rainfall amount still retains a strong negative association with daily water consumption. Over the period, the correlation coefficient is -0.34; while over the summer, it becomes stronger and the value reaches to -0.43. As indicated in Figure 3.4, the rainfall occurs over all seasons. Even in the winter, there are some heavy rain days; however, these rainfalls may have no impact on outdoor usage. Therefore, summer rainfalls maintain stronger relations with daily water consumption than the rest of the year. Thus, daily rainfall amount and rainfall frequency information get included in the initial weather feature set.

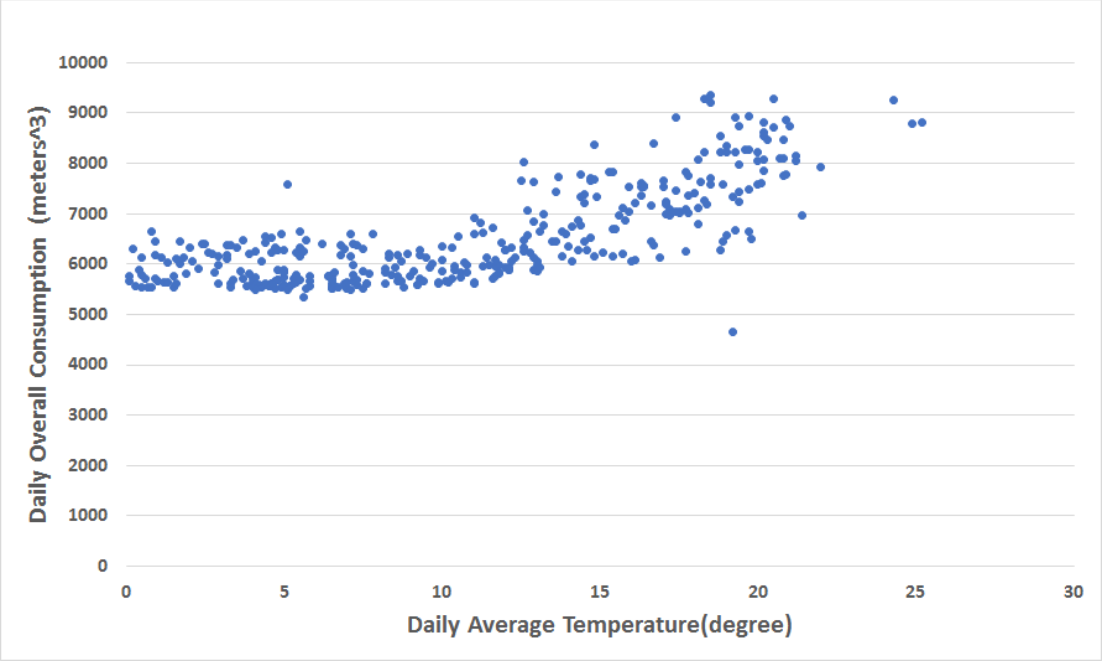


Figure 3.3: Association between daily water consumption and average daily temperature at the dissemination area level.

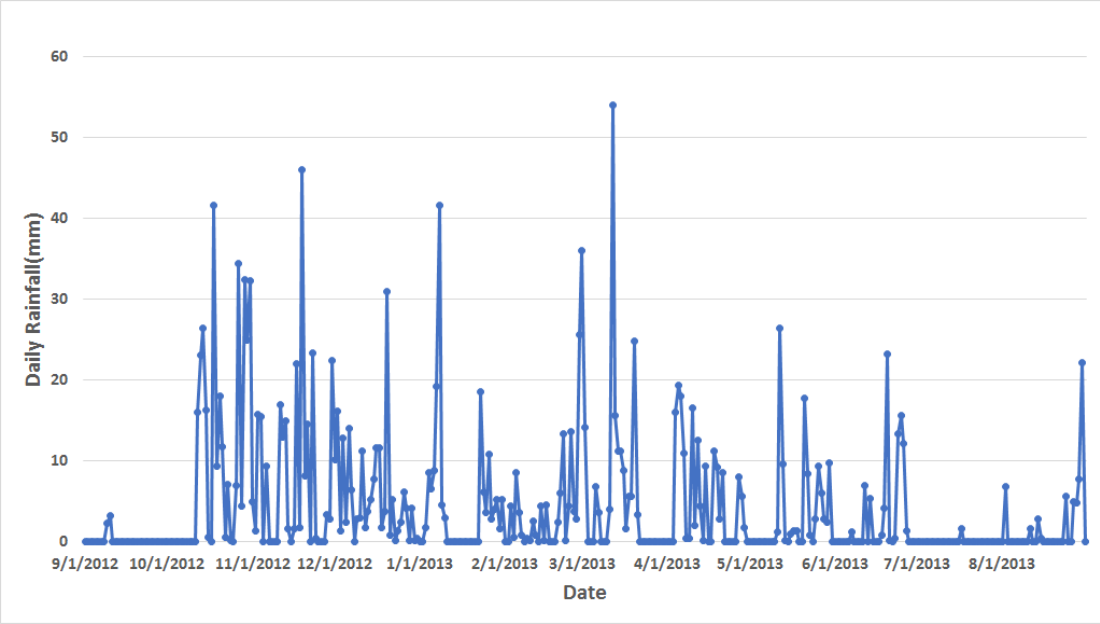


Figure 3.4: Daily rainfall amount over the period September 2012 to August 2013.

3.4 Demographic Data

Demographic information is collected from Statistics Canada. However, the information is assembled at the dissemination area level for all residents, not only single-family residents. One assumption for this feature set is that the averaged demographic information well represents the statistics of the single-family residences. In the initial raw feature set, there are over 300 variables; however, this research leverages prerequisite knowledge (see Chapter 2) and compresses the list to twelve features. In order to demonstrate associations between demographic features and water consumption, average hourly water consumption over the entire period are aggregated in dissemination levels. The associations are evaluated by the correlation coefficient between each feature and the consumption value. The results are presented in Table 3.2.

Over all variables, education features have the strongest relationship with water consumption. EDU_Level1 retains a correlation of 0.52, which implies that the higher the percentage of people in a dissemination area that acquired a post-secondary diploma, the higher the average water consumption that area requires. As well-educated people are normally at high income positions, their water consumption is insensitive to water pricing and their amount of usage is higher than that of others. On the other hand, the

Table 3.2: Demographic features at the dissemination area level and the correlation between the features and hourly water consumption.

Feature	Correl.	Comment
EmployE	−0.31	Employment rate in the area
EmployU	0.31	Unemployment rate in the area
TaxBelow	0.11	Tax return below average (Percentage)
TaxAbove	−0.11	Tax return above average (Percentage)
BDROccupied1	0.01	Percentage of 1 bedroom occupied in this area
BDROccupied2	0.09	Percentage of 2 bedroom occupied in this area
BDROccupied3	−0.34	Percentage of 3 bedroom occupied in this are
BDROccupied4	0.19	Percentage of 4 bedroom occupied in this area
EDULevel1	0.52	Percentage of people acquired post-secondary diploma
EDULevel2	0.04	Percentage of people acquired high-school diploma
EDULevel3	−0.46	none of level1 or level2
TaxPerPerson	−0.35	Median tax per person

employment rate for adults and average water consumption have a negative association with a coefficient of −0.31. Employees spend around 7.5 hours at work and there is no water consumption during this period. The more time employed people are at work, the less time they stay at home and the less water they consume. Somewhat unexpected is that the median tax per person has a negative correlation with water consumption. The median tax per person for all dissemination areas and the corresponding average hourly consumption are plotted in Figure 3.5. The plot suggests that when the tax amount is lower than $24,000, water consumption has significant variance. The lower value implies that there could be more part-time workers or retired people in the area. These people normally spend more time at home as compared to other employees; therefore, they consume more water. Lastly, the percentage of three bedrooms in a dissemination area outperforms other bedroom occupation percentages and show a close relationship with hourly water consumption. The negative relationship suggests that residents living in 3-bedroom houses are the most conservative clients.

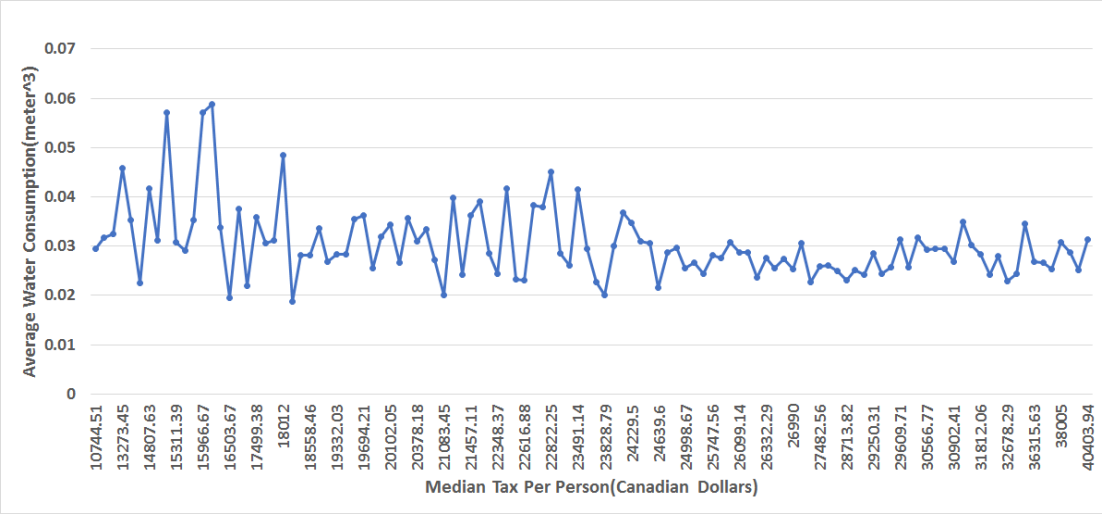


Figure 3.5: Median tax per person vs water consumption at the dissemination area level.

3.5 Assessment Data

Housing property information is commonly used for residential water consumption prediction. This research engages property information from the British Columbia Assessment Authority. Out of thirteen features in the raw data, six features are chosen and listed in Table 3.3. Original property information is at the household level; however, this research focuses on predictions at the dissemination area level. Hence, the raw data is aggregated and values are preserved for the corresponding attributes. The correlation coefficient values between each variable and the average water consumption are calculated to demonstrate associations. The averaged housing area variable retains the strongest association with water consumption. Larger properties imply larger families and larger backyards. These factors contribute to more water consumption. An unexpected correlation is that the average building year of an area and water consumption retain a positive correlation. Normally newly built properties are considered more water efficient; however, the positive correlation indicates that newer house residents eventually consume more water than others. This abnormal behavior can be explained by the financial situations of residents. People in newer properties may be in better financial situations as newer properties are sold at higher values than older ones. Wealthy people usually consume more water in gardening and irrigation.

Table 3.3: Aggregated property assessment features at the dissemination area level and the correlation between the aggregated features and hourly water consumption. The features are aggregated by taking the averages of the individual household information.

Feature	Correl.	Comment
AvgHouseArea	0.54	Property total areas in squared feet.
AvgStories	0.13	Number of stories of the property
AvgYearBuilt	0.49	Building year of the property
AvgBeds	0.15	Number of bedrooms in a property
AvgFullBath	0.36	Number of full bathrooms in a property
AvgPartBath	−0.13	Number of partial bathrooms in a property

3.6 Summary

In this chapter, I first describe data preparation procedures for this research. Since this research focuses on predictions at the dissemination area level, how household level data gets aggregated is demonstrated. Furthermore, I conduct data analysis on all feature sets including weather, demographic and property assessment information. Correlation coefficient values are leveraged to measure associations between each individual variable and the average hourly water consumption. At last, I identify and analyze pivotal variables and build persuasive datasets that are utilized in feature selection and model construction.

Chapter 4

Feature Selection

In this chapter, I first demonstrate what feature selection is and highlight previous work. Next, I illustrate feature selection algorithms used in this thesis. Since I consider that feature selection is problem and data dependent, I conduct feature selection for all different scenarios: one hour ahead with and without preselected features for weekday and weekend consumption predictions, and one hour ahead features for weekday and weekend usage forecasts. Lastly, a proposed feature set is determined for each scenario which will be used in the model construction and selection process (see Chapter 5).

4.1 Feature Selection in General

Feature selection is essential for building predictive models. Especially for high dimensional datasets, the task of determining which features should be used is important. For example, Yu and Liu [56] consider that high dimensioned datasets may degrade performance of learning algorithms due to irrelevant and redundant features. Hence, optimizing feature sets by removing redundant and irrelevant features conserves model computation time and improves model accuracy. Feature selection is considered one of the most crucial steps for this research. Therefore, it is necessary to examine how a feature set is built for each experiment before demonstrating model construction and selection.

There are two approaches to feature selection: the wrapper approach and the filter approach. The wrapper approach leverages prediction models to score feature subsets, whereas the filter approach utilizes a proxy measure to score feature subsets.

In each step of the wrapper approach, there are two types of features: selected features and candidate features. Following the feature selection step, a selection method assesses each candidate feature by adding it to the selected feature set, then builds a training model, evaluates the model and adds a feature that improves the model accuracy the most at that stage. For example, let $F_s = \{f_1, f_2, ..., f_k\}$ be a selected feature set from the previous step, and let $F_c = \{f_m, ..., f_z\}$ be the corresponding candidate feature set. The approach then requires to build models for $F_{t_m} = \{f_1, f_2, ..., f_k, f_m\}$, ..., $F_{t_z} = \{f_1, f_2, ..., f_k, f_z\}$. Finally, the method picks F_{t_i} from F_{t_m}, ..., F_{t_z} that builds a model with the best accuracy. Feature selections for this process are computational intensive, especially when there are a large number of features.

The filter approach requires much less computation. Here a method ranks features based on various criteria such as mutual information or a correlation coefficient. However, a generated feature set, which consists of the top k number of features in the ranking list, may not be as effective as the one generated with the wrapper method.

4.2 Feature Selection for Predicting Water Usage

Although computationally more intensive, I prefer to use the wrapper approach due to its improved accuracy. In order to downgrade computational complexities, each feature selection in this project is divided into two steps. The first step is selecting hourly water usage variables. The water consumption data in this research is time series data. Clarifying the relationships between the target hourly usage and preceding hourly consumption enriches the relevant candidate feature set. For example, previous one hour, two hours and a week before the same hour information is selected by Herrera and [29] for hourly water usage predictive models. The second step is merging features selected in step one with date (e.g., what day of the week it is and whether it is a public holiday), daily weather (e.g., daily average wind speed, temperature and rainfall), demographic (e.g., education level and income at the dissemination area level) and house property information (e.g., average number of bedrooms per property and building year of the property) and designing an overall optimized feature set.

Different experiments are defined in table Table 4.1. Experiments are performed using two types of scenarios: one hour ahead and one day ahead hourly water usage predictions. For each scenario, two subcategories, weekday and weekend models, are developed except the one hour ahead baseline model (see Chapter 5). To further investigate feature selection contributions, one hour ahead models with feature selection are further divided into "with" and "without" preselected features models. The fundamental difference between with and

Table 4.1: Definitions of different experiment scenarios.

Experiment Scenario	Definition
One hour ahead	Predict the water consumption of next hour.
One day ahead	Predict the water consumption of the same hour in tomorrow.
Baseline model	The model tackles weekday and weekend predictions together and all the features fed into the model are only based on human knowledge without any feature selection algorithms.
With preselected features	Engage some features based on human knowledge in the initial feature set before applying feature selection algorithms.
Without preselected features	The feature selection process is in the form of no human in the loop and all the features are selected by algorithms.

without preselected feature experiments is whether there is a human in the loop process. In the with preselected feature experiments, human knowledge is engaged initially and used to build a seed feature set. In contrast, there are no human engagements in the without preselected feature experiments and all the features are selected based on feature selection algorithms. In order to choose effective features, feature selection is conducted for each experiment, which ensures the best model and feature compatibility. Feature selection results are described below; they demonstrate that different models eventually prefer different features.

Before demonstrating how features are selected, it is important to present the data used to build feature sets. There are 111 dissemination areas eventually selected (See Chapter 3) and each dissemination area contains 8,760 time series data about user consumption as well as features from weather, demographic, date and property datasets. Engaging all data into feature selection is infeasible because of the very high computational time. Therefore, a prerequisite of feature selection is targeting a subset of data that successfully represents the entire experiment data. Since there are different populations in different dissemination areas, all these areas are divided into 10 subgroups based on the population. One dissemination area is randomly selected in each subgroup to prevent biased selections of certain groups.

The data for ten selected dissemination areas is combined and divided into two groups: weekend and weekday datasets. There are four experiments implemented for one hour ahead models: pre-selected features on weekday and weekend datasets, without preselected

features on weekday and weekend datasets, and two experiments implemented for one day ahead models: weekday and weekend datasets. For each scenario, feature selection is conducted five times. Since selected historical water consumption features will be combined with features from other datasets, the number of features selected for this group is limited to 10. It prevents final feature selection from relying only on historical water consumption data and avoiding features from other datasets.

4.3 Relevant Preceding Hourly Usage Selection

Previous work demonstrates that preceding time series data is pivotal in predicting future time series values. Altunkaynak and Ozger [7] present a fuzzy method for predicting future monthly water consumption values from three antecedent water consumption. Zahrani and Monasar [5] develop a forecasting model to predict daily water consumption by coupling time series models and ANNs. The models consider effects of three past daily water consumption values and climatic variables. However, how much delay should be tracked is data dependent. This research discovers interdependencies within time series data. Since one hour and one day ahead predictions leverage different hourly consumption data, feature selection for these two criteria is demonstrated separately in the following context.

There are 48 predictive variables engaged in previous hourly water consumption experiments. In order to better present previous hour usage, the notation t_i is used in the rest of this section, where t_i represents the the water usage in the ith previous hour. For example, $t_1$68 means the water usage in the hour 168 hours previous to the current hour. MATLAB's `sequentialfs`, a forward wrapper method, is chosen for this research (see Table 4.2 for the parameters to the function). The method selects a subset of features from a given dataset that provides the best predictions for a target variable by sequentially selecting features until no more features can be added to improve the accuracy. This algorithm takes a function that returns a criterion measuring distances between predicted values and actual values for the testing dataset. In this research, a linear model is selected and the mean root square error is calculated as the performance measurement. For each experiment, feature selection is reprocessed five times to ensure the stability of the result. Moreover, 10-fold cross validation is applied to avoid overfitting for each feature selection process.

Table 4.2: Parameters to MATLAB's `sequentialfs` algorithm for feature selection.

Name	Value	Comments
fun	linear model	Used RMSE performance measurement
x	input matrix	Contains 8,569 instances with 48 columns
y	prediction vector	Contains 8,569 hourly water usage values
cv	cross validation	10-fold cross validation was chosen
keepin	features that must be include in final selection	Previous 1, 2, 3 and 168-hour water usage were selected for the pre-defined models.

4.3.1 One Hour Ahead Historical Water Consumption Feature Selection

For one hour ahead water usage prediction, two sets of previous data are recruited. One set consists of the previous 24 hours of the target hour, which are from t_1 to t_{24}. Another set consists of the same hour a week before to 23 hours ahead, which are from t_{168} to t_{191} hours. There are 48 predictive variables in an initial feature set. As mentioned above, two different feature sets are built, one with preselected features and one without preselected features. For preselected feature sets, t_1, t_2, t_3 and t_{168} are preselected. All experimental results are listed in Table 4.3 and Table 4.4.

Table 4.3: Selected historical water consumption features for the scenario of *with* preselected features; i.e., t_1, t_2, t_3 and t_{168} are preselected.

weekdays	t_1	t_2	t_3	t_{168}	t_{169}	t_{23}	t_{170}	t_{24}	t_{10}	t_{178}
weekends	t_1	t_2	t_3	t_{168}	t_{169}	t_{23}	t_{170}	t_{172}	t_{22}	t_{20}

Table 4.4: Selected historical water consumption features for the scenario of *without* preselected features.

weekdays	t_{168}	t_1	t_{169}	t_{23}	t_{170}	t_3	t_{24}	t_{10}	t_{178}	t_{173}
weekends	t_{168}	t_1	t_{169}	t_{23}	t_{170}	t_3	t_{22}	t_{20}	t_{191}	t_{190}

In pre-selection experiments, results demonstrate that in addition to the pre-selected four features, there are three features that are commonly selected: t_{23}, t_{169} and t_{170}. Feature t_{23} can be considered as a proxy for one-hour ahead water usage information in the future. Combined with t_{168} information, the information for three consecutive hours around the target hour in the previous week are collected. Hence a weekly consumption pattern is suggested. Two other important features in weekday models of pre-selections are t_{10} and t_{178}. Feature t_{178} is 10 hours ahead of last week the same hour. Figure 4.1 illustrates consecutive 168 hours water usage from Sunday 12:00am to Saturday 23:00pm. The figure suggests that there are two peak periods in each weekday. One is in the morning and the other is in the evening. There is a 10-hour difference between the evening and morning peak hours. After each peak period, hourly water usage drops in different patterns. For example, on Monday, after the morning peak hour (point 32), it takes around 7 hours for total usage to drop from 388 to 300 cubic meters; however, after the evening peak hour (point 43), total hourly usage drops from around 416 to below 65 cubic meters in around 7 hours. Although both usage groups drop after peak hours, one drops rapidly and the other smoothly. In this situation, t_{10} becomes very useful. For example, 10-hour usage before morning peak hour is always significantly lower than current hour usage. On the other hand, 10-hour usage before evening peak hour is either very close to or higher than current hour usage. In the without preselection experiments, the selections agree on eight out of ten features with preselection results for weekend data. It adds t_{191} and t_{190} hours usage information at the last two steps. For weekdays, instead of choosing t_2 information, the non-preselection approach chooses t_{30} information at the last step. Experiments suggest that t_2 information may not be useful for both weekdays and weekends. However, a weekly pattern is strongly indicated. Features t_{168}, t_{169} and t_{170} are selected in high priorities for all criteria. Moreover, a daily pattern is inferred. Feature t_{23} is one of the top selected features in all the scenarios and the feature t_{24} is selected for weekdays in both experiments.

4.3.2 One Day Ahead Historical Water Consumption Feature Selection

Instead of using the previous 24 hours, which are not available for one-day-ahead prediction, and the consecutive 24 hours before the same hour last week, this research leverages t_{24} hour to t_{168}, which are from the previous 24 hours to the same hour last week. Although there are 145 features initially, eventually 10 features are selected. Hence the computational complexity of the wrapper approach is still manageable, and the same algorithms and procedures that are used in the one hour advanced feature selection are applied. At the end, two sets of features are built, one for weekdays and one for the weekend. Stepwise

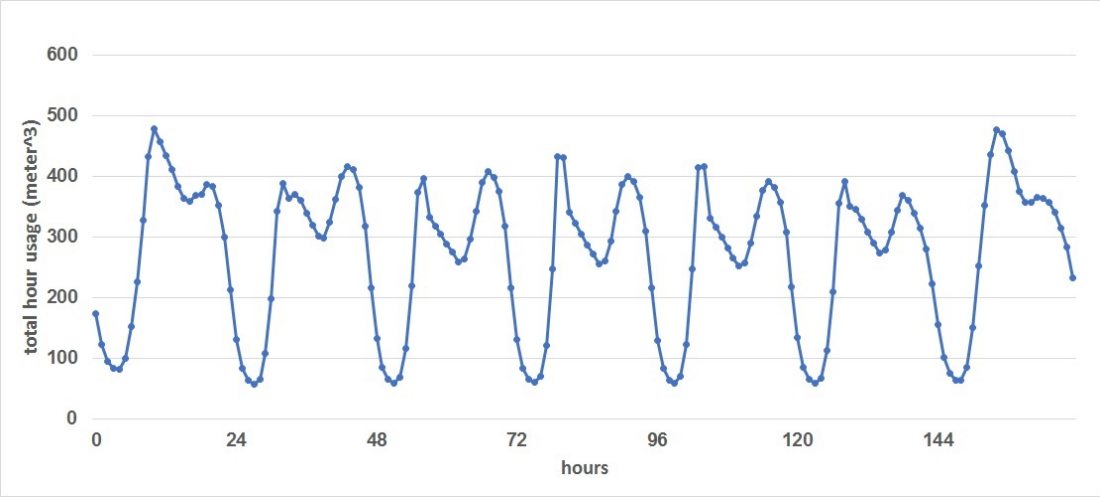


Figure 4.1: Hourly water usage from Sunday to Saturday.

feature selection results for weekdays and weekends are listed in Table 4.5. Ten previous hourly usage predictive variables are included in each feature set.

Table 4.5: Selected historical water consumption features for one day head scenario.

weekdays	t_{24}	t_{144}	t_{168}	t_{48}	t_{95}	t_{120}	t_{34}	t_{35}	t_{72}	t_{96}
weekends	t_{24}	t_{72}	t_{168}	t_{153}	t_{157}	t_{74}	t_{36}	t_{113}	t_{155}	t_{27}

For weekdays, the same hour water consumption information from previous one day to seven days are included. This suggests that weekday hourly usage follows a daily pattern. In other words, hourly usage has significant dependencies on the same hour usage during the entire week. By contrast, weekend hourly usage does not show remarkable dependencies; however, the previous one day and previous week the same hour information is selected. This confirms the result of the one hour in advance experiment and demonstrates an hourly pattern and a weekly pattern for hourly usage.

4.4 Combining All Features

One of the contributions of this research is engaging various datasets including demographic, weather and assessment information for prediction. All ten historical water con-

sumption features determined from previous subsection and twenty-eight additional features from distributed datasets constitute an initial feature set. The same dissemination area date is applied to this step. In contrast to previous subsections, there are no limitations on the number of features in the resulting feature set.

Witten and Frank [55] define REPTree as a fast decision tree. It uses information gain or variance and is pruned by leveraging reduced-error pruning. The combined feature selection of this research uses the REPTree algorithm implemented in the Weka application. Feature selection follows the wrapper approach and 10-fold cross validation is leveraged to reduce risks of overfitting. For REPTree configurations, the minimum number of instances in each leaf is set to 16 and there are no limitations on the depth of the tree. For each experiment, the number of times each feature is selected is recorded.

4.4.1 One Hour Ahead Final Feature Selection

The resulting feature set for one hour ahead models are presented in Table 4.6. Most features selected in the final feature set are from the historical water usage dataset. This confirms the presumption that hourly water usage data is highly correlated, and it is a pivotal predictor for hourly consumption prediction models. Another commonly selected feature is the day of week information. During weekdays, although water usage patterns are identical on Tuesdays, Wednesdays, and Thursdays, Monday and Friday patterns are slightly different from others. For example, on Friday peak hours come one hour earlier than on other weekdays. The peak starts from 17:00, while for the other weekdays it starts from 18:00 as shown in Figure 3.1. This may be because people get home earlier on Friday night and start to enjoy their weekends. Saturday and Sunday consumption patterns are not only significantly different from weekdays, but they are also different from each other. On Saturdays, evening peak hours are not obvious. The amount of water consumption smoothly drops down from 17:00 to the end of the day. However, on Sundays, hourly usage increases after 18:00, reach a peak value and then drop down.

From the weather feature set, the most commonly selected features are daily average temperature and rainfall information. This conforms previous work results. For example, in summer time, most water consumption is the result of outdoor usage, including filling swimming pools, irrigation and car washing. Outdoor consumption is determined mainly by these two weather variables. For filling swimming pools, the higher the temperature is, the better the chance that residents leverage the facility and fill their pools. For irrigation, when experiencing many consecutive days without rainfalls, residents are more likely to irrigate their front and back yards.

Table 4.6: The feature selection results of all one hour ahead scenarios.

	Weekdays		Weekends	
Type	Non-preselection	Preselection	Non-preselection	Preselection
Water	t_1	t_1	t_1	t_1
	t_{10}	t_{10}	t_{168}	t_{168}
	t_{168}	t_{168}	t_{169}	t_{169}
	t_{169}	t_{169}	t_{170}	t_{170}
	t_{170}	t_{170}	t_{191}	
	t_{173}			t_2
	t_{178}	t_{178}		t_{20}
	t_{23}	t_{23}	t_{22}	t_{22}
	t_{24}		t_{23}	t_{23}
	t_3	t_3	t_3	t_3
		t_2		
Weather	temperature	temperature	temperature	temperature
	previousRainfall	previousRainfall	previousRainfall	
	rainfall	rainfall	rainfall	rainfall
	sinceLastRainFall	sinceLastRainFall	windspeed	windspeed
	previousLastRainFall			barometer
Date	is_holiday	is_holiday		
	week_day		week_day	week_day
Property	AvgBeds	AvgBeds	AvgBeds	AvgBeds
			AvgYearBuilt	AvgYearBuilt
			AvgStories	AvgStories
		AvgHouseArea	AvgHouseArea	AvgHouseArea
	AvgFullBath	AvgFullBath		AvgFullBath
Demographic	Tax_Per_Person	Tax_Per_Person	Tax_Per_Person	Tax_Per_Person
	EDU_Level1	EDU_Level1	EDU_Level1	EDU_Level1
	Tax_B			Tax_B
		EDU_Level3	EDU_Level3	
			Employ_E	

Regarding the property feature set, average bedrooms per single family residence is included in all four experiments. This factor can be used to explain base water usage, and the number of residents in a family can be inferred from it. The more people live in a property, the more water consumption it requires.

The last feature set contains demographic information. Tax per person information and first level of education are commonly included in all experiments. Financial situations have been proven to be an effective factor for water predictions (see Chapter 2). Families with better financial situations normally consume more water.

4.4.2 One Day Ahead Final Feature Selection

The result feature set for one day ahead models is shown in Table 4.7. Not surprisingly, ten previous hourly usage variables are included in the final feature sets. It confirms once again the importance of characteristics in historical hourly usage information. For weekdays, the same hour water consumption information from previous one day to seven days are included. This suggests a daily pattern, which confirms one hour ahead experimental results. By contrast, weekend hourly usage does not show significant daily patterns; however, previous one day and previous week the same hour information is selected. This implies a one day ahead dependency and a weekly dependency for weekend models.

It is astonishing that the weekday and weekend resulting feature sets highly agree on property and weather information. All variables from the weather dataset are selected. The importance of the temperature factor is underscored in this experiment. It is selected ten out of ten times for both weekday and weekend models. However, there are different considerations for rainfall information. For weekdays, since last rainfall (the number of days since last rainfall) is selected seven out of ten times; while rainfall (the actual amount of rainfall on the next day) is selected only one out of ten times. By contrast, rainfall is selected nine out of ten times, whereas since last rainfall is selected only three out of ten times for weekend feature selection. This illustrates the dispute about which of these two variables is more important (see Chapter 2). One achievement of this research is that it provides persuasive reasons for when one feature dominates the other by leveraging effective feature selection algorithms. From the assessment feature set, the number of bathrooms, which retains close association with water consumption as an individual variable (see Chapter 3), is ignored for both weekday and weekend feature sets. It is another contribution of feature selection that it eliminates features that are considered pivotal in regard to individuals but irrelevant or redundant while considering them with others in a group.

Table 4.7: The feature selection results of all one day ahead scenarios.

Type	Weekdays	Weekends
Water	t_{24}	t_{24}
	t_{34}	t_{27}
	t_{35}	t_{36}
	t_{48}	t_{72}
	t_{72}	t_{74}
	t_{95}	t_{113}
	t_{96}	t_{153}
	t_{120}	t_{155}
	t_{144}	t_{157}
	t_{168}	t_{168}
Weather	Temperature	Temperature
	SinceLastRainFall	SinceLastRainFall
	Rainfall	Rainfall
	Barometer	Barometer
	Windspeed	Windspeed
	PreviousLastRainFall	PreviousLastRainFall
	PreviousRainfall	PreviousRainfall
Date	week_day	week_day
	is_holiday	
Property	AvgBeds	AvgBeds
	AvgHouseArea	AvgHouseArea
	AvgYearBuilt	AvgYearBuilt
	AvgStories	AvgStories
Demographic	Tax_Per_Person	Tax_Per_Person
	BDR_Occupied_1	BDR_Occupied_1
	BDR_Occupied_2	BDR_Occupied_2
	BDR_Occupied_3	BDR_Occupied_3
	BDR_Occupied_4	BDR_Occupied_4
	Tax_B	Tax_B
	Employ_E	Employ_E
	EDU_Level1	
	EDU_Level2	
	EDU_Level3	

4.5 Summary

In this chapter, I demonstrate detailed feature selection for each scenario. A two-step feature selection is conducted: first, select ten previous water consumption features, then determine effective features from 10 historical consumption, weather, demographic, date and property datasets. As a result, different features are selected for each criterion. As emphasized, features for a particular model is data dependent and customized feature selection processes can benefit model construction and improve prediction accuracy.

Chapter 5

Model Alternatives and Model Selection

In this chapter, I leverage accomplishments from the data preparation and feature selection steps to implement models for hourly consumption predictions of one hour and one day in advance scenarios at the dissemination area level. This chapter addresses the following hypotheses.

Hypotheses:

1. Separate weekend and weekday models can improve model performance.
2. Feature selection can assist models to achieve better accuracy compared to ones constructed without feature selection.
3. Multilayer models can outperform single hidden layer models with proper feature sets.
4. With proper model selection, one day ahead models can perform as accurately as one hour ahead models.

Regarding the first two hypotheses, three main types of models are built: one hour ahead baseline models, one hour ahead refined models with preselected features, and one hour ahead refined models without preselected features. The latter two types will be called refined models in the rest of the thesis. Baseline models are implemented by using features from previous research work and tackle weekday and weekend predictions in one single

model. By contrast, refined models use corresponding feature sets built in Chapter 4 and separate weekday and weekend predictions into two models. To investigate the third hypothesis, one day ahead models are implemented. Models are built based on three scenarios: single hidden-layer models, two hidden-layer models and three hidden-layer models. All one day ahead models share the same input feature set. Lastly, proposed models representing one hour and one day in advance predictions are compared to verify the last hypothesis.

All ANNs in this research are implemented with the MATLAB `fitnet` routine. A challenge for model implementation is acquiring appropriate parameter settings. Parameters to be configured are listed in Table 5.1. This research selects three candidate initial learning rates: 0.1, 0.01 and 0.001. As experiments show that models with 0.001 outperform others, only 0.001 models are discussed in the rest of the thesis. To ensure that the optimal number of nodes in the hidden layer(s) is determined, this research experiments on ANNs with several options. For one hidden-layer models, the number of nodes in the hidden layer ranges from four to fifteen. For two hidden-layer models, the number of first hidden layer nodes ranges from five to thirteen, and the second hidden layer retains the number of nodes from three to the number of nodes in the first hidden layer minus one. For three hidden-layer models, the first hidden layer retains five to eleven nodes, the second hidden layer maintains three to the number of node in the first hidden layer nodes minus one nodes, and the third hidden layer retains two to the number of nodes in the second hidden layer minus one nodes. Models are evaluated using 10-fold cross validation. Due to the limitations of Matlab 2013a, the fitnet function is not able to obtain the mean or sum absolute mean error as the evaluation function during its training process. Therefore, the metrics mean squared error is selected for the fitnet function instead.

Table 5.1: Model configurations for all ANN models in the experiment section.

Configuration	Values	Comments
Initial learning rate	0.1, 0.01, 0.001	Optimal value is 0.001
Nodes		Different models retain different nodes for experiment
Training algorithm	Trainlm	
Evaluation function	Mean squared error	
Routine	`fitnet`	
Data separation	Training, validation and testing	Training dataset contains 10% of the data; validation and testing contain 90%

At the end, the predicted values for all individual hours of all dissemination areas are collected and used in model evaluation. Mean absolute error (AE) and mean absolute percentage error (APE) are used to measure prediction accuracy.

- Mean absolute error (AE): $\frac{1}{n}\sum_{t=1}^{n}|PredictValue_t - ActualValue_t|$
- Mean absolute percentage error (APE): $\frac{100}{n}\sum_{t=1}^{n}\left|\frac{PredictValue_t - ActualValue_t}{ActualValue_t}\right|$

Since models are going to predict average single family hourly consumption in each dissemination area and average hourly consumption are in low volumes (even peak hour usage are normally just around 400 liters), thresholds for acceptable model performance differences are set to 0.020 liters per hour and 0.05 percent for AE and APE respectively. Models with performance in these ranges are considered comparable; otherwise, the model with better performance is considered to dominate the other. Although the thresholds are set to these relatively small digits, the statistics are in per single family resident per hour level. When adding the entire dissemination area single resident consumption all together, the number turns to be significant and could impact the utility company decisions. In addition to tackling tie breakers, different quantiles of AE and APE are calculated. From a model structure perspective, the number of connections (or weights) in a model is considered as an evaluation factor. The threshold for this factor is set to 50. The rules of thumb for model evaluation and model selection for this research are listed below.

Rules:

1. The first step is eliminating the overfitted models.
2. The second step is filtering models by using thresholds of AE and APE.
3. The third step is further obsoleting model candidates by leveraging thresholds of model complexity, as measured by the number of connections in the network.
4. If necessary, the last step is using tie breakers (different quantiles of absolute error and absolute percentage error) to make a final decision.

As indicated above, numerous models are implemented in this research. I evaluate models from two aspects: model performance as measured by prediction accuracy and model complexity as measured by the number of connections. As suggest by Coelho and Richert

[49], the costs of prediction loss are not the same for all instances. Only the instances with high hourly water consumption are considered during performance evaluations. For weekdays, hours of high water consumption are 7:00–9:00am and 6:00–9:00pm. For weekends, the hours are from 8:00am to 10:00pm.

5.1 One Hour Ahead Models

In this section, I present models for hourly consumption prediction of one hour in advance at the dissemination area level. The following models are presented: a baseline model, a single hidden layer weekend model, and a single hidden layer weekday model.

5.1.1 Baseline model

Two distinct characteristics of the baseline models are no feature selection and a single model that predicts both weekdays and weekends. Instead of applying feature selection, important features from previous work are collected.

Herrera and Torgob [29] suggest that in order to predict next hour water usage, crucial historical water consumption information and climate information should be included. These include previous one hour consumption, previous two hours consumption, last week the same hour consumption, temperature, wind velocity, atmospheric pressure and rainfall. Adamowski and Chan [1] implement a daily water prediction model selecting features from historical water consumption including up to four days ahead water usage data and weather information, such as maximum daily temperature and daily total precipitation. Besides historical water consumption and weather information, Agthe and Billings [3] indicate that property characteristics can explain variations in water demand as well. Selected property features include the number of bedrooms and property age. After transferring these properties to data available in this research, the feature set includes average number of bedrooms, average property age and average property area in each dissemination area.

Many studies consider that social economic information can explain water consumption. Junguo [42] studies house income, house size and water price impacts on water usage. Since there are no direct applicable variables from this research dataset, the possible replacement variables are tax per person, percentages of families' income tax above average, and percentages of families' income tax below average.

To respect previous work achievements on variables that explain water consumption,

this research leverages previous one hour, previous two hours, last week the same hour combined with variables in Appendix 1 as inputs to the baseline ANNs.

Another characteristic of the baseline model is that it predicts all weekday and weekend consumption in a single model. As weekdays and weekends follow significantly different water consumption patterns in all dissemination areas, the day of week information is engaged to assist the model to distinguish different usage behaviors.

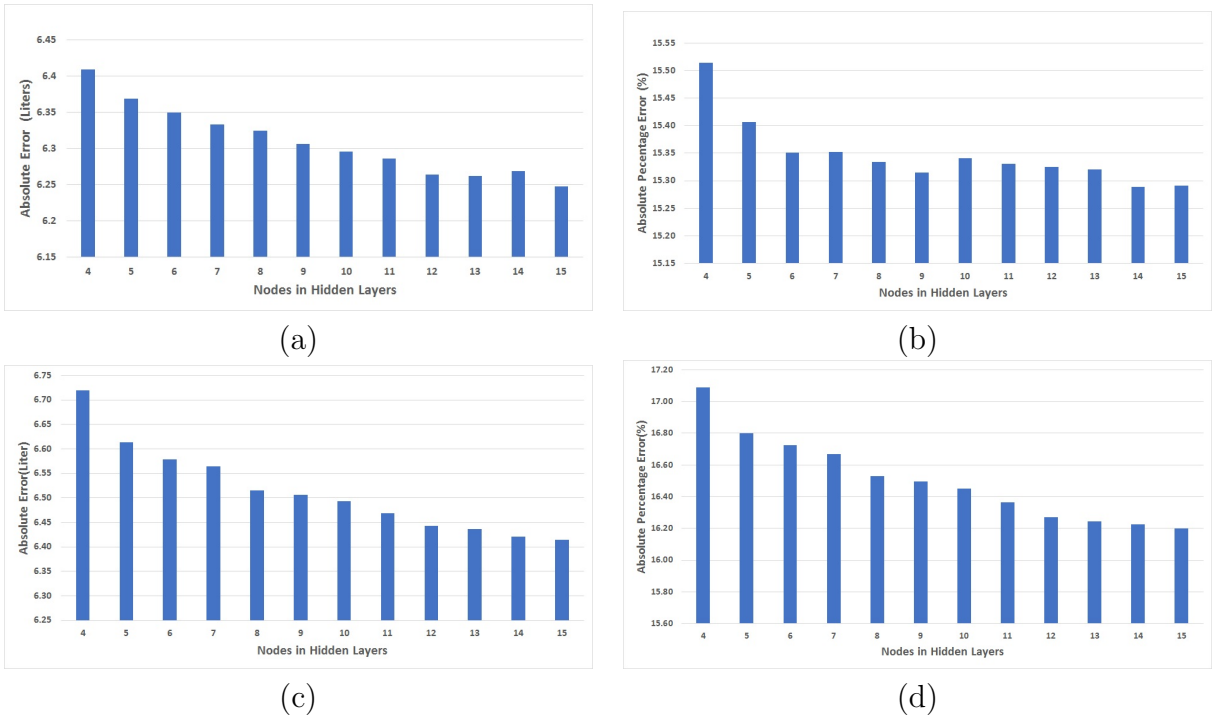


Figure 5.1: Performance of one hour ahead baseline model on *weekend* peak hours as measured by (a) absolute error and (b) absolute percentage error, and *weekday* peak hours as measured by (c) absolute errors and (d) absolute percentage errors. Note: y-axis does not start at zero.

For weekend peak hours, from an AE perspective in Figure 5.1(a), model performance monotonically increases until the number of nodes in the hidden layer reaches 12. There is a significant increase from 11 nodes to 12 nodes in the hidden layer; however, after 12, improvements become unstable. From an APE perspective in Figure 5.1(b), performance monotonically improves until the number of nodes in the hidden layer reaches 9 and then improvements become unstable. As with weekend peak hours, performance statistics of weekday peak hours are presented in Figure 5.1(c) and Figure 5.1(d) for AE and APE values

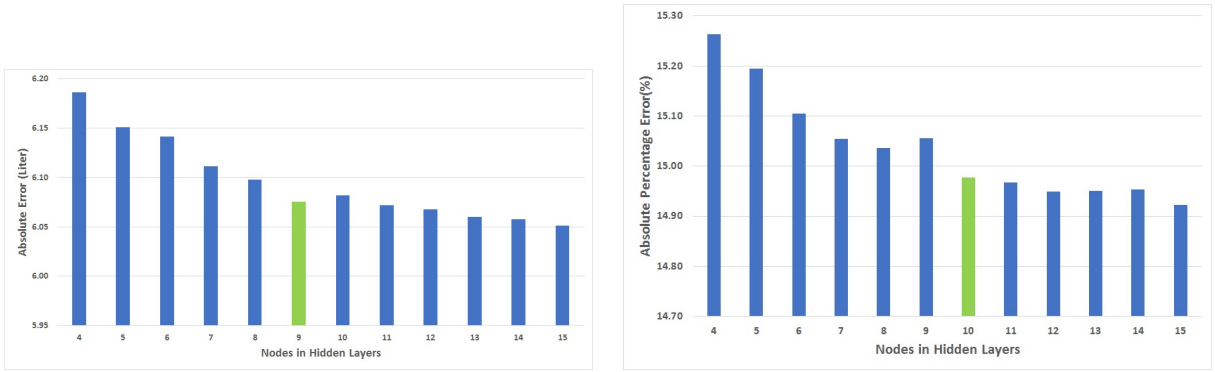
respectively. Model performance keeps increasing as additional nodes are included in the hidden layers. However, the improvement in AE and APE values slows after the number of nodes reaches 12 in the hidden layer. Using the rules for model selection established above, the model with 12 hidden nodes is chosen and considered as the baseline model. All models with less than 12 hidden layer nodes are not comparable to model-12 as performance differences are over the thresholds. From the standpoint of model complexity, model-12 has significantly fewer connections than models with more nodes in the hidden layers but with negligible or no decrease in accuracy. Therefore model-12 is considered as the best choice of model for one hour ahead baseline weekday predictions.

5.1.2 Hour Ahead Single Hidden Layer Weekend Model

In this subsection, one hour ahead weekend models are presented. Weekend models are divided into with and without preselected feature scenarios.

Weekend models without preselected features.

Average errors (AE) and average percentage errors (APE) for all experiments are presented in Figure 5.2(a)&(b), respectively. As mentioned above, although all hourly usage are predicted, only peak hours' results are used in evaluations.



(a) (b)

Figure 5.2: Performance of one hour ahead without preselected features on *weekend* peak hours measured by (a) absolute error and (b) absolute percentage error. Note: y-axis does not start at zero.

Table 5.2: Without preselected features *weekend* performance (AE and APE) and model complexity during peak hours.

Nodes	4	5	6	7	8	9	10	11	12	13	14	15
AE (10^{-2} liter)	618.6	615.1	614.2	611.2	609.8	607.5	608.2	607.2	606.8	606.0	605.8	605.1
APE (0.1 percent)	152.7	151.8	151.6	150.9	150.5	150.3	150.3	150.1	15.00	149.9	149.8	149.6
Connections	93	116	139	162	185	208	231	254	277	300	323	346

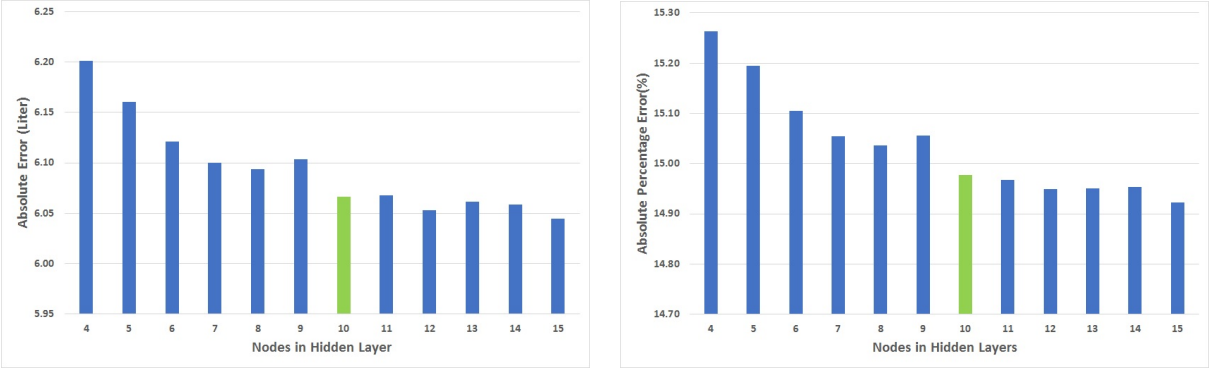
From an AE perspective, Figure 5.2(a) indicates that after the number of nodes in hidden layers reaches 9, model performance shows only small changes for models with additional nodes in the hidden layer. To further investigate this proposed model the number of connections are compared in Table 5.2. Each entry represents a model with the corresponding number of nodes in its hidden layer. Model-9 is considered as the best choice of model. By using a predefined threshold 0.02, models with fewer nodes in their hidden layers are eliminated. In the same table, APE statistics indicate that proposed model candidates are from model-8 to model-15 by following the same comparisons in AE, but with a threshold 0.05%. Thereafter, models to be considered are narrowed down to model-9 to model-14. The last step is model complexity comparisons. From the model complexity perspective, models with more than 11 nodes in hidden layers exceed the complexity threshold while comparing with model-9; therefore, they are eliminated. Model-10 is rendered obsolete as it not only requires 23 additional connections, but also retains worse peak hour AE as compared to model-9. Model-11 requires 46 additional connections, which is close to the complexity threshold; however, performance improvements are notable. Therefore, the selected model for an hour ahead weekend without preselected features is model-9.

Weekend models with preselected features.

Average errors (AE) and average percentage errors (APE) for all experiments are presented in Figure 5.3(a)&(b), respectively.

Table 5.3: Preselected features *weekend* performance (AE and APE) and model complexity during peak hours.

Nodes	4	5	6	7	8	9	10	11	12	13	14	15
AE (10^{-2} liter)	620.1	616.1	612.1	610.0	609.4	610.4	606.7	606.8	605.3	606.1	605.9	604.5
APE (0.1 percent)	152.6	152.0	151.1	150.6	150.4	150.6	150.0	150.0	149.5	149.5	149.5	149.2
Connections	97	121	145	169	193	217	241	265	289	313	337	361



(a) (b)

Figure 5.3: Performance of one hour ahead with preselected features on *weekend* peak hours measured by (a) absolute error and (b) absolute percentage error. Note: y-axis does not start at zero.

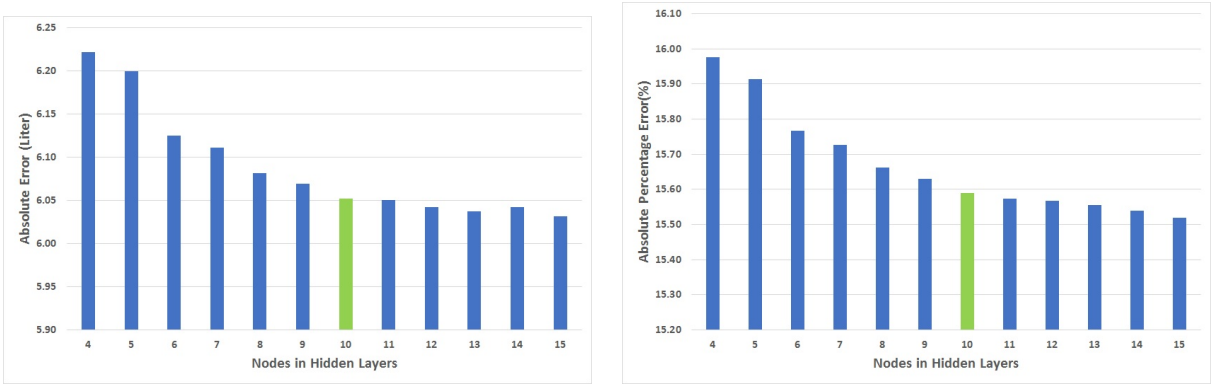
From an AE perspective, model performance goes up and down after the number of nodes in hidden layers reaches 10. Comparing models' AE statistics in Table 5.3, only model-15 preserves better performance than model-12; model-10 and model-11 are also recruited, since they retain performance differences within the threshold range. Hence a candidate list generated from an AE perspective contains model-10, model-11, model-12 and model-15. From an APE perspective, model-12 is also considered as a benchmark. By considering APE statistics shown in the same table and a threshold value of 0.05%, model candidates from model-10 to model-15 are considered. Lastly, model complexity comparisons consider that model-10 outperforms all others as it has at least 24 fewer connections, but still being only negligibly less accurate. Therefore, the model selected for one hour ahead weekend with preselected features is model-10.

5.1.3 Hour Ahead Single Hidden Layer Weekday Model

In this subsection, one hour ahead weekday models are presented. Models are separated into with and without preselected feature scenarios, and one optimal model is determined for each criterion.

Weekday models with preselected features.

The same as weekend models, Figure 5.4 and Figure 5.4(b) demonstrate overall model performance measured by AE and APE statistics. From both an AE and an APE perspective, model performance increases monotonically until the number of nodes in hidden layers reaches 13. From Table 5.4, acceptable candidates include models from model-10 to model-15. Lastly, rules of model complexity suggest that model-10 is a reasonable selection because it requires significantly fewer connections (shown in Table 5.4) while maintaining performance within an acceptable level.



(a) (b)

Figure 5.4: Performance of one hour ahead with preselected features on *weekday* peak hours as measured by (a) absolute error and (b) absolute percentage error. Note: y-axis does not start at zero.

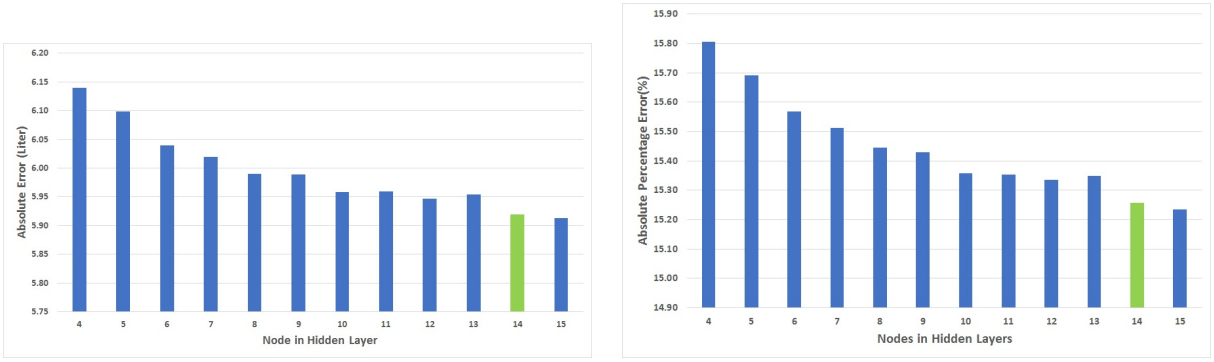
Table 5.4: Preselection models *weekday* performance (AE and APE) and model complexity during peak hours.

Nodes	4	5	6	7	8	9	10	11	12	13	14	15
AE (10^{-2} liter)	622.2	620.0	612.5	611.1	608.2	607.0	605.2	605.0	604.2	603.7	604.2	603.2
APE (0.1 percent)	159.8	159.1	157.7	157.3	156.6	156.3	155.9	155.7	155.7	155.6	155.4	155.2
Connection	93	116	139	162	185	208	231	254	277	300	323	346

Weekday models without preselected features.

As with other experiments, model performance significantly increases initially as the number of nodes in hidden layers increases from both an AE and an APE perspective, as

demonstrated in Figure 5.5 and Figure 5.5(b) respectively. However, once the number of nodes in hidden layers reaches 10, performance adjustments are negligible until it reaches 14, which retains significant performance improvements. As performance statistics indicate in Table 5.5, the only model comparable to model-14 is model-15, since model-14 dominates all others over thresholds from a performance perspective. Lastly, the table demonstrates that model-15's complexity increases approximately half way to the complexity threshold. Since the performance difference from AE statistics is negligible between model-14 and model-15, model-14 is considered to be the best selection.



(a) (b)

Figure 5.5: Performance of one hour ahead without preselected features on *weekday* peak hours measured by (a) absolute error and (b) absolute percentage error. Note: y-axis does not start at zero.

Table 5.5: Without preselected features *weekday* performance (AE and APE) and model complexity during peak hours.

Nodes	4	5	6	7	8	9	10	11	12	13	14	15
AE (10^{-2} liter) 614.0	609.9	603.9	601.9	599.0	598.9	595.9	596.0	594.7	595.4	592.0	591.3	
APE (0.1 percent)	158.1	156.9	155.7	155.1	154.5	154.3	153.6	153.6	153.4	153.5	152.6	152.3
Connections	97	121	145	169	193	217	241	265	289	313	337	361

5.2 One Day Ahead Models

In this section, one day in advance models are presented and compared. In contrast to one hour ahead models, models in this subsection are implemented not only in a single hidden

layer, but also in multiple hidden layers. There are six sets of models including the ones with single, two and three hidden layers for both weekend and weekday scenarios.

5.2.1 Day Ahead Weekend Model

Table 5.6: Single hidden layer a day ahead *weekend* models performance (AE and APE) and model complexity during peak hours.

Nodes	4	5	6	7	8	9	10	11	12	**13**	14	15
AE (10^{-2} liter)	698.3	693.4	690.6	686.9	684.3	682.1	679.7	676.6	674.5	**673.0**	671.7	670.2
APE (0.1 percent)	172.9	171.9	171.2	170.4	169.7	169.3	168.6	167.9	167.3	**166.9**	166.6	166,2
connections	125	156	187	218	249	280	311	342	373	**404**	435	466

In this subsection, one day in advance weekend models for one, two and three hidden layers are introduced first. For the single hidden-layer scenario, there are 12 models implemented with hidden layer nodes from 4 to 15. For two hidden-layer models, the number of nodes in first hidden-layer starts from 5 to 13, and the number of nodes in second hidden-layer starts from 3 to the number of nodes in the first hidden layer minus 1. Hence, there are 54 models generated. Finally, three hidden-layer models are introduced. The number of first hidden layer nodes is between 5 and 11, the number of second layer nodes is between 3 and the number of first layer nodes minus 1, and the third layer nodes start from 2 to the number of nodes in the second layer minus 1. Therefore, there are 119 models generated in total. At the end, one proposed model is selected for each criterion. As the reasoning behind the selection of each of the models is similar to that already detailed above for one-hour-ahead models, fewer details are presented.

Table 5.6 shows the performance and number of connections for a range of possible single hidden-layer models. Model-13 is chosen as the best tradeoff of performance and complexity.

Table 5.7 shows the performance and number of connections for a range of possible two hidden-layer models. Model-12-4, an ANN with 12 nodes in the first hidden layer and four nodes in the second hidden layer, is selected to be the best model representing one day ahead weekend two hidden-layer models, as it dominates other models on the performance with only a small increase in model complexity.

Table 5.8 shows the performance and number of connections for a range of possible three hidden-layer models. Model-11-7-2, an ANN with 11 nodes in the first hidden layer, seven nodes in the second layer, and two nodes in the third hidden layer, is selected to be the best model representing one day ahead weekend three hidden-layer models.

Table 5.7: Two hidden layer a day ahead *weekend* models performance (AE and APE) and model complexity during peak hours.

Connections	First	Second	AE (10^{-2} liter)	APE (0.1 percent)
337	10	3	676.6	167.9
349	10	4	676.5	168.2
361	10	5	674.9	167.7
370	11	3	673.7	167.6
373	10	6	673.5	167.4
383	11	4	675.0	167.6
385	10	7	676.0	168.1
396	11	5	675.3	167.8
397	10	8	674.6	167.8
403	12	3	672.7	167.1
409	10	9	673.0	167.2
409	11	6	674.1	167.4
417	**12**	**4**	**670.3**	**166.6**
422	11	7	673.7	167.5

5.2.2 Day Ahead Weekday Model

In this subsection, one day in advanced weekday models are implemented and compared. By following the same procedure as with weekend models, there are 12, 54 and 119 models for single, two and three hidden-layer scenarios respectively. At the end of each subsection a representative model is selected. As the reasoning behind the selection of each of the models is similar to that already detailed above for one-hour-ahead models, fewer details are presented.

Table 5.9 shows the performance and number of connections for a range of possible single hidden-layer models. Model-13 is chosen as the best tradeoff of performance and complexity.

Table 5.10 shows the performance and number of connections for a range of possible two hidden-layer models. Model-10-5, an ANN with 10 nodes in the first hidden layer and five nodes in the second hidden layer, is selected to be the best model representing one day ahead weekday two hidden-layer models.

Table 5.11 shows the performance and number of connections for a range of possible

Table 5.8: Three hidden layer a day ahead *weekend* models performance (AE and APE) and model complexity during peak hours.

AE (10^{-2} liter)					APE (0.1 percent)				
Connection	First	Second	Third	Perf.	Connection	First	Second	Third	Perf.
386	9	7	5	675.0	377	10	5	3	167.7
391	9	8	4	675.0	386	9	7	5	167.7
396	10	7	2	673.1	396	10	7	2	167.4
419	11	6	2	672.8	427	11	6	3	167.1
422	10	9	2	673.0	**433**	**11**	**7**	**2**	**166.9**
427	11	6	3	672.0	435	11	6	4	167.3
433	**11**	**7**	**2**	**670.8**	442	11	7	3	167.1
442	11	7	3	672.6	443	11	6	5	166.8
443	11	6	5	670.8	447	11	8	2	167.4

three hidden-layer models. Model-11-7-3, an ANN with 11 nodes in the first hidden layer, seven nodes in the second layer, and three nodes in the third hidden layer, is selected to be the best model representing one day ahead weekday three hidden-layer models.

Table 5.9: Single hidden layer a day ahead *weekday* models performance (AE and APE) and model complexity during peak hours.

Nodes	4	5	6	7	8	9	10	11	12	**13**	14	15
AE (10^{-2} liter)	674.3	666.8	664.4	659.6	655.1	651.5	651.6	648.4	646.0	**642.7**	643.7	640.8
APE (0.1 percent)	172.1	170.2	169.8	168.5	167.6	166.6	166.6	165.9	165.3	164.6	**164.9**	164.0
Connections	141	176	211	246	281	316	351	386	421	**456**	491	526

5.3 Summary

This chapter discusses the implementation of three sets of models: one hour ahead baseline models, one hour ahead refined models, and one day ahead refined models. Detailed model selections based on well-specified criteria are presented. There are five models proposed for one hour ahead predictions: one for baseline models and one for each preselected feature models and without preselected feature models of weekdays and weekends. Another six models are proposed for one day ahead hourly water consumption predictions: one for each

Table 5.10: Two hidden layer a day ahead *weekday* models performance (AE and APE) and model complexity during peak hours.

Connections	First	Second	AE (10^{-2} liter)	APE (0.1 percent)
389	10	4	647.2	165.8
395	9	8	647.2	165.6
401	**10**	**5**	**644.2**	**164.9**
413	10	6	645.8	165.2
414	11	3	646.0	165.3
425	10	7	646.7	165.6
427	11	4	643.7	164.7
437	10	8	644.5	165.1
440	11	5	642.9	164.5
449	10	9	643.4	164.6
451	12	3	643.4	164.8
453	11	6	642.9	164.6
465	12	4	643.5	164.8
466	11	7	643.0	164.6
479	11	8	642.8	164.5
479	12	5	643.7	164.7
492	11	9	642.3	164.3

single hidden layer models, two hidden layers models, and three hidden layers models of weekdays and weekends. In the next chapter, I will use the models proposed in this chapter to validate the four hypotheses mentioned at the beginning of this chapter.

Table 5.11: Three hidden layer a day ahead *weekday* models performance (AE) and model complexity during peak hours.

connections	first	second	third	AE (10^{-2} liter)
486	**11**	**7**	**3**	**641.4**
519	11	10	2	641.2

Chapter 6

Model Evaluation and Discussion

In this chapter, I conduct holistic comparisons among twelve proposed models from Chapter 5. By analyzing the strengths and weaknesses of each model, I indicate common inaccurately and well-forecasted instances. Lastly, I argue that investigations of model performance make contributions in understanding the challenges of water predictions and I provide suggestions about how to overcome these difficulties.

6.1 Experimental Results

In this section I compare twelve proposed models from Chapter 5. Comparisons are divided into two subsections, one hour ahead and one day ahead model comparisons. Also, a representative model of each subsection is collated to conclude performance differences between one hour and one day ahead models.

6.1.1 One Hour Ahead Models Comparisons

The first set of comparisons in this section are on one hour ahead models. Three models are engaged: a baseline model, a model without preselected features, and a model with preselected features. Weekday and weekend model performance collations are separated to further analyze model distinctions.

One hour ahead weekday models.

Statistics for the proposed models for one hour ahead hourly water consumption on weekdays are listed in Table 6.1. Models implemented by following feature selection procedures significantly dominate the baseline model from the perspectives of both model complexities and prediction accuracy. The result strongly supports the first two hypotheses: separated models for weekdays and weekends with appropriate feature selection notably improve model performance. However, when comparing models with and without preselected features, conclusions are disputable. Figure 6.1 and Figure 6.2 indicate that the model without preselected features outperforms the other from both an AE and an APE perspective for all levels; moreover, the advantages in AE and APE are above error thresholds. By contrast, the model with preselected features is simpler having 95 fewer connections than the one without preselected features. The major costs of model complexities only occur at model construction time since once a model is built, input data is just fed into the model and effects of complexity costs are negligible. Hence model accuracy are the most important consideration. The best model without preselected features, Model-14, is determined to be a proposed model for this group.

Table 6.1: Comparison of proposed one hour ahead weekday models.

	baseline	without preselection	with preselection
connections	397	337	242
nodes	12	14	10
AE (10^{-2} liter)	644.2	592.0	605.2
APE (0.1 percent)	162.7	152.6	155.9

One hour ahead weekend models.

AE and APE statistics of three proposed models are listed in Table 6.2. In addition to the number of connections, the number of nodes in each hidden layer and the average statistic values are presented. The baseline model is dominated by others from all levels. The baseline model not only requires the most number of connections, but also has significant performance disadvantages. Models with feature selection and weekend and weekday separations gain at least 130 connections from a model complexity perspective; moreover, they retain the advantages in average AE at a level of 10^{-1} liters, which is far beyond the

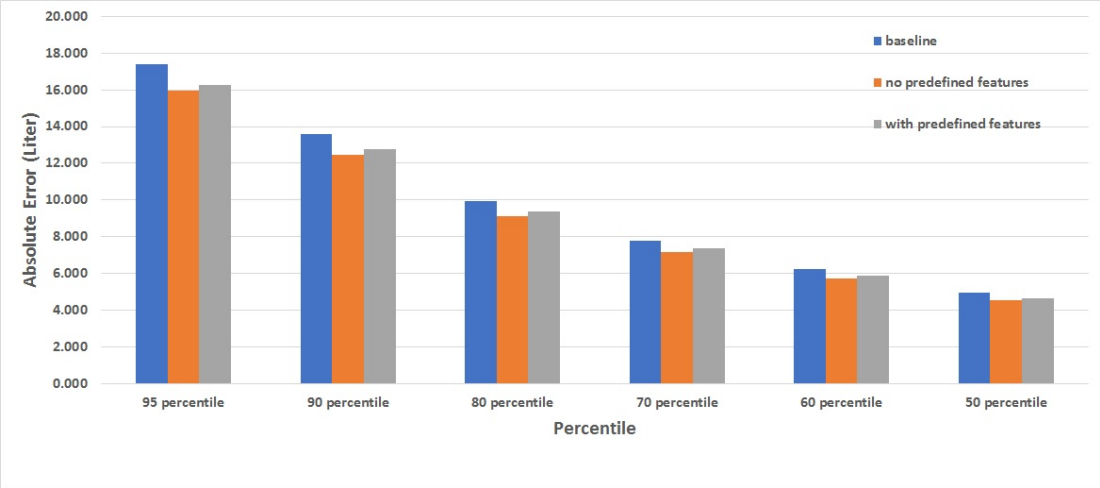


Figure 6.1: One hour ahead overall models comparisons weekdays as measured by absolute error.

threshold 0.02 liters. Similarly, APE advantages are close to or over 0.3% on average. Furthermore, 95 percentiles to 50 percentiles (medium value) of AE and APE statistic values are presented in Figure 6.3 and Figure 6.4. The results show the improvement given by of feature selection and separate weekday and weekend models. For models with and without preselected features, decisions are disputable. From a model complexity perspective, the without preselected model retains 33 connections less, which is significant but within an acceptable range. From an AE perspective, although performance differences are negligible in some statistics, the ones for 95 percentiles and 60 percentiles are significant and are over the threshold. Performance differences are more obvious from an APE perspective. Differences of all performance statistics are over the threshold 0.05%. Above all, the one with preselected features outperforms the one implemented purely by using the feature selection. Although it diverges from my hypothesis, it illustrates the importance of human experiences during feature selection. One conclusion of this research is that prerequisite knowledge of a study area combined with proper feature selection algorithms can improve model accuracy. In the end, the best model with preselected features, model-10, is selected to present models for one hour ahead weekend predictions.

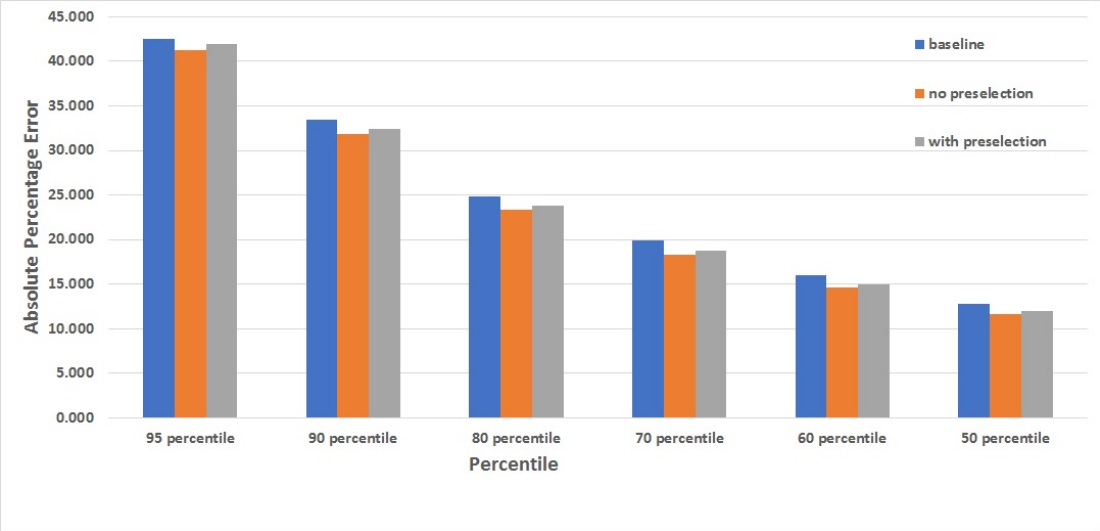


Figure 6.2: One hour ahead overall models comparisons weekdays as measured by absolute percentage error.

Table 6.2: One hour ahead proposed weekend models comparisons in AE and APE.

	baseline	without preselection	with preselection
connections	373	208	241
nodes	12	9	10
avg(10^{-2} liter)	626.4	607.5	606.7
avg(0.1 percent)	153.3	150.3	149.8

6.1.2 One Day Ahead Models Comparisons

One day ahead weekday models.

AE and APE statistics of one day ahead weekday proposed models are presented in Table 6.3. From a model complexity perspective, the model with two hidden-layer retains significant number of connections less than others; therefore, it is considered as a baseline model. From average an AE and an APE perspective, the baseline model performance is notably worse than the other two. Figure 6.5 and Figure 6.6 indicate that single hidden-layer and three hidden-layer models retain significant advantages in performance as compared to the baseline model at high percentiles; however, differences converge as it goes

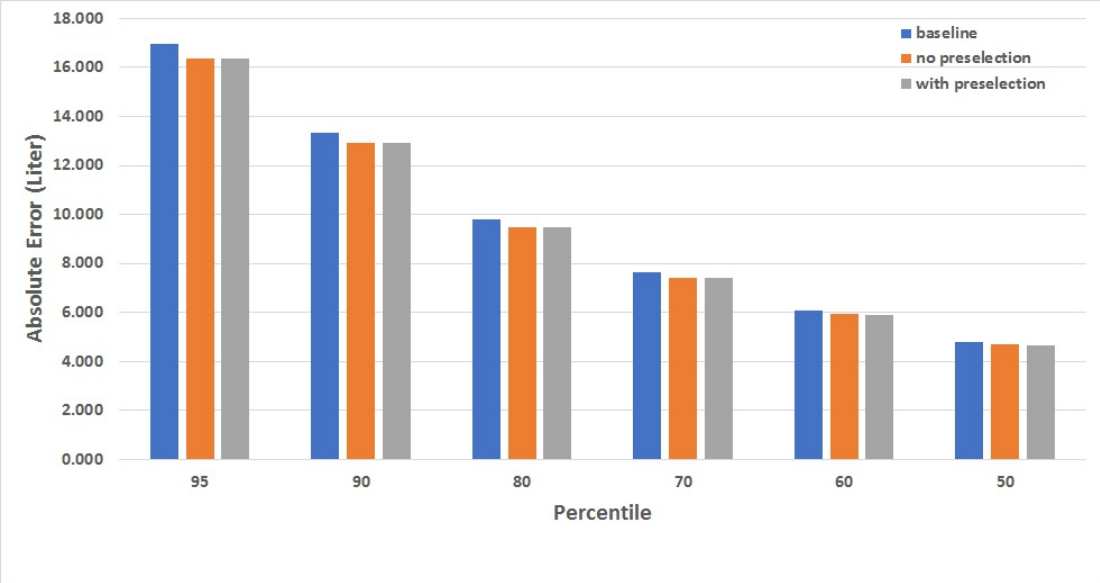


Figure 6.3: One Hour Ahead Overall Models Comparisons Weekends AE.

from 95 percentiles to 50 percentiles. Both average AE and APE of the baseline model are worse than the proposed three hidden layers model, model-11-7-3, by 2.9×10^{-2} and $6.8 \times 10^{-2}\%$ respectively, which are around 1.5 times the threshold values; however, the 85 connection advantages make it 1.7 times less than complexity thresholds. Therefore, the baseline model outperforms the three hidden-layer model. By contrast, although performance disadvantages persist for the baseline model as compared to the single hidden-layer model, performance differences are within acceptable ranges. However, more than 50 connections advantages in model structures make the baseline model a better choice. Hence, the proposed two hidden layers model, model-10-5, is determined to present one day ahead weekday predictions.

One day ahead weekend models.

AE and APE statistics of three models are presented in Table 6.4. In contrast to weekday models, model complexity differences are not significant. The best two hidden layers model, model-12-4, is selected as the benchmark model during comparisons. From average an AE perspective, the two hidden-layer model outperforms the proposed single hidden layer model, model-13, by 2.7×10^{-2} liters per hour, which exceeds the threshold; however, 95 to 50 percentiles of AE in Figure 6.7 suggest that the single hidden-layer model dominates the

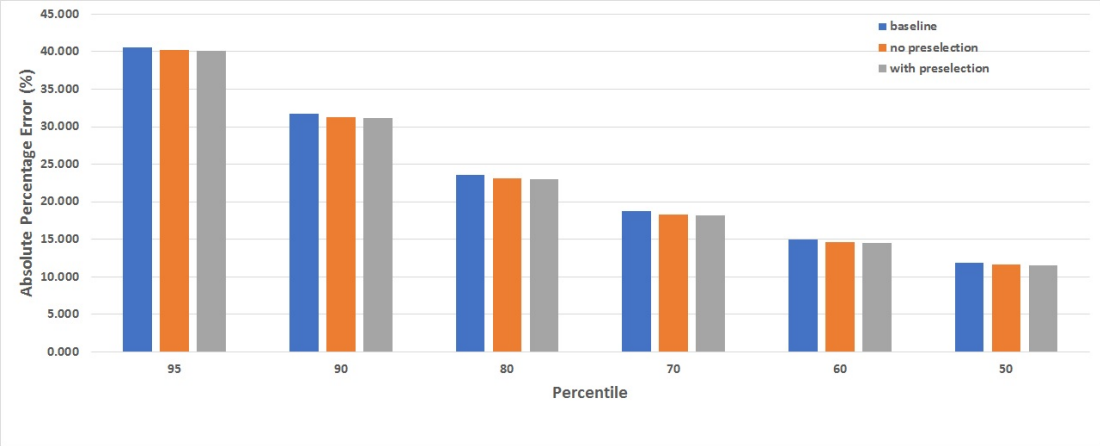


Figure 6.4: One Hour Ahead Overall Models Comparisons Weekends APE.

Table 6.3: One day ahead proposed weekday models comparisons in AE and APE.

hidden layer(s)	single	two	three
connections	456	401	486
third hidden layer nodes			11
second hidden layer nodes		10	7
first hidden layer nodes	13	5	3
AE (10^{-2} liter)	642.7	644.2	641.3
APE 0.1 percent	164.6	164.9	164.2

benchmark model. This implies that the benchmark model retains remarkable accuracy rates on half of instances. From an APE perspective, the two hidden-layer model gains at least $3.0 \times 10^{-2}\%$ over the other two on average errors. Although APE statistics from 95 percentiles to medium values in Figure 6.8 suggest that the single hidden-layer model dominates the benchmark one, the average advantages strongly imply that the benchmark model retains remarkably accurate rates on half of the instances predictions. Hence, the benchmark model outperforms the single hidden-layer model. On the other hand, the benchmark model not only retains better performance than the three hidden-layer model, but also requires less connections. Thus, the best two hidden layer model, model-12-4, is considered to be the one representing one day ahead weekend models.

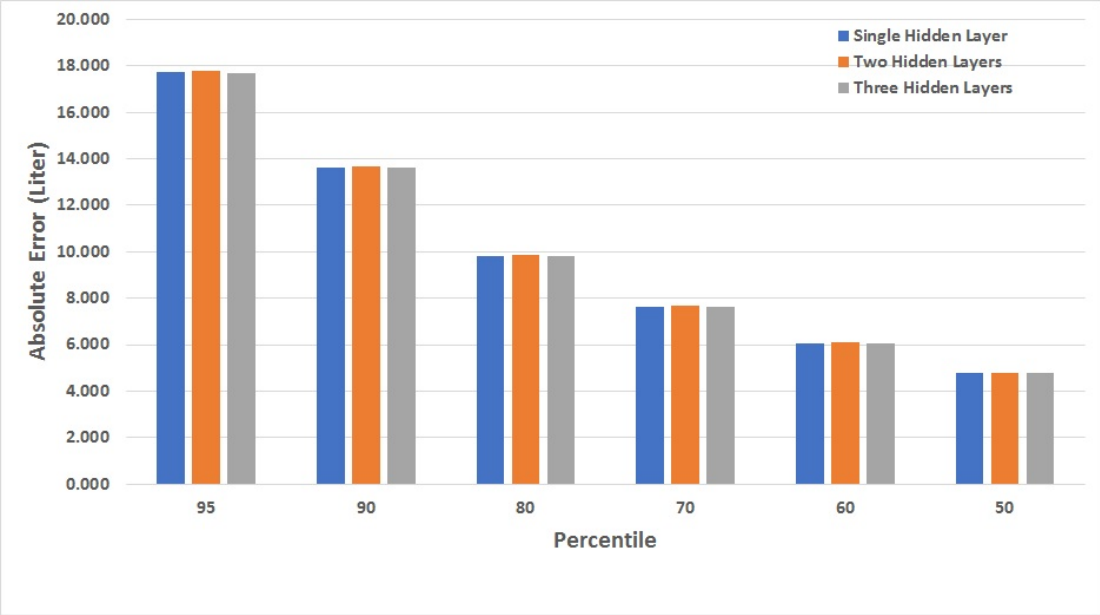


Figure 6.5: One Day Ahead Overall Models Comparisons Weekdays AE.

6.2 Discussion

After determining four proposed models, in the following subsections I analyze the results in detail, and confirm or reject the hypotheses from Chapter 5.

First, it is pivotal to confirm the improvements offered by the ANN models. Baseline experiments are conducted. One day ahead same hour and one week ahead same hour information is directly applied to represent the predicted hourly consumption for each scenario. Moreover, the comparisons are tackled by splitting the weekday and weekend consumption. By following the same evaluation processes of other experiments, only peak hours predictions are investigated. The statistics of each experiment are demonstrated in Table 6.5. As the statistics of each experiment suggested in Table 6.5, the base experiments are dominated by baseline models, one hour ahead models and one day ahead models implemented in this research. Therefore, the results confirm the importance of engaging ANN models.

From one hour ahead experiments, results indicate that the baseline model not only gives significantly lower performance than others, but also requires more connections. From these experiments, the first conclusion is that models with feature selection and differentiations between weekdays and weekends give higher accuracy when predicting hourly water

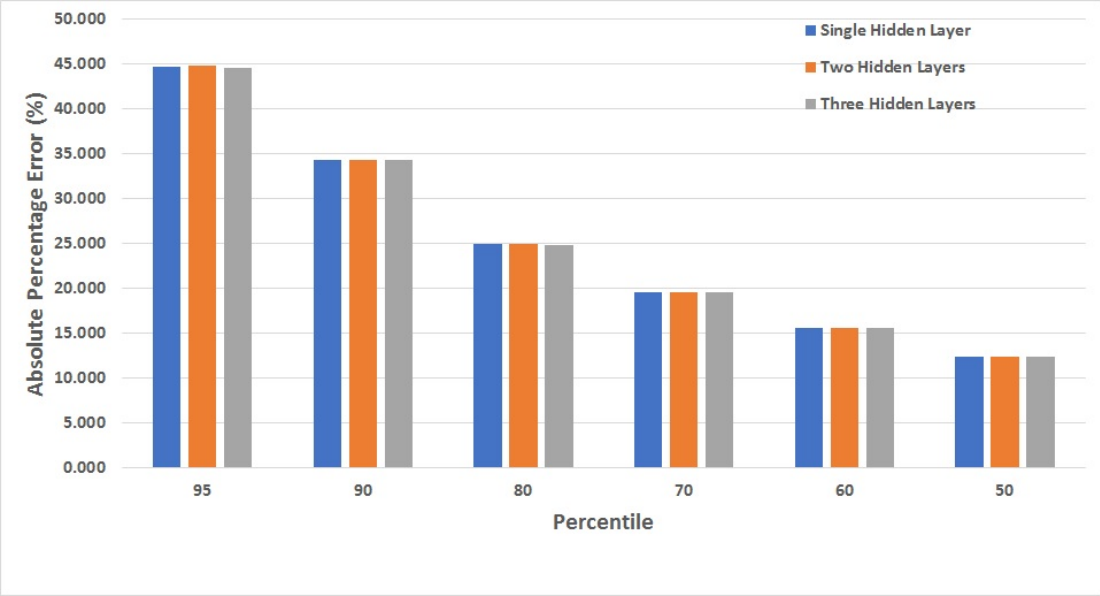


Figure 6.6: One Day Ahead Overall Models Comparisons Weekdays APE.

consumption. Further to feature selection, weekday and weekend experiments suggest different conclusions. Weekday results show that although the model without preselected selections requires more connections than the one with preselected features, performance improvements are impressive. AE and APE differences are six times higher than threshold levels. Hence it demonstrates the advantages of relying on feature selection algorithms over human intuition. By contrast, weekend statistics illustrate that the model with preselected features outperforms the one with feature selection only. Although the without preselected feature model maintains less connections and negligible AE differences, the notable APE differences make great contributions on determining the proposed model. Hence, whether feature selection should engage human inputs or only rely on algorithms is an empirical question and remains open to debate.

For one day ahead models, comparisons are conducted among one, two and three hidden layer models. In contrast to one hour ahead results, weekday and weekend results highly agree with each other. The two hidden-layer model dominates the other two. This confirms the hypothesis that multi hidden-layer models can achieve or even overcome single hidden-layer model performance with identical model structures.

The last set of experiments consists in conducting comparisons between one hour ahead and one day ahead models. AE and APE statistics of four proposed models and two

Table 6.4: One day ahead proposed weekend models comparisons in AE and APE.

hidden layer(s)	single	two	three
connections	404	417	433
third hidden layer nodes			11
second hidden layer nodes		12	7
first hidden layer nodes	13	4	2
AE (10^{-2} liter)	673.0	670.3	670.8
APE (0.1 percent)	166.9	166.6	166.9

Table 6.5: Model performance of using previous day or previous week information to predict water consumption in AE and APE.

	AE (10^{-2} liter)	APE (0.1 percent)
one day ahead weekdays	892.9	226.3
one week ahead weekdays	885.8	227.4
one day ahead weekends	1028.2	241.7
one week ahead weekends	876.3	214.8

baseline models are presented in Figure 6.9 and Figure 6.10 respectively. Average errors and 95 percentiles to median errors are presented to provide a holistic picture of all models' performance. These figures clearly demonstrate that one hour ahead models dominate one day ahead models in all levels from both an AE and an APE perspective. It implies that short-term water consumption data contributes more than longer term water consumption data for hourly water predictions. Hence, the more recent accurate data can be collected from smart meter devices, the more accurate model predictions can be provided to utility companies.

In order to further investigate the model performance, the metrics of overestimated, underestimated and precisely estimated values are demonstrated in Table 6.6. The absolute error threshold value utilized during model comparisons is leveraged in this process. The "overestimated above threshold" field contains the value representing the percentage of overestimated forecasts predicting 0.02 liters more than the actual usage. In contrast, the "underestimated above threshold" field represents the percentage of underestimated instances which the actual consumption is 0.02 liters than the predicted values. Over all the

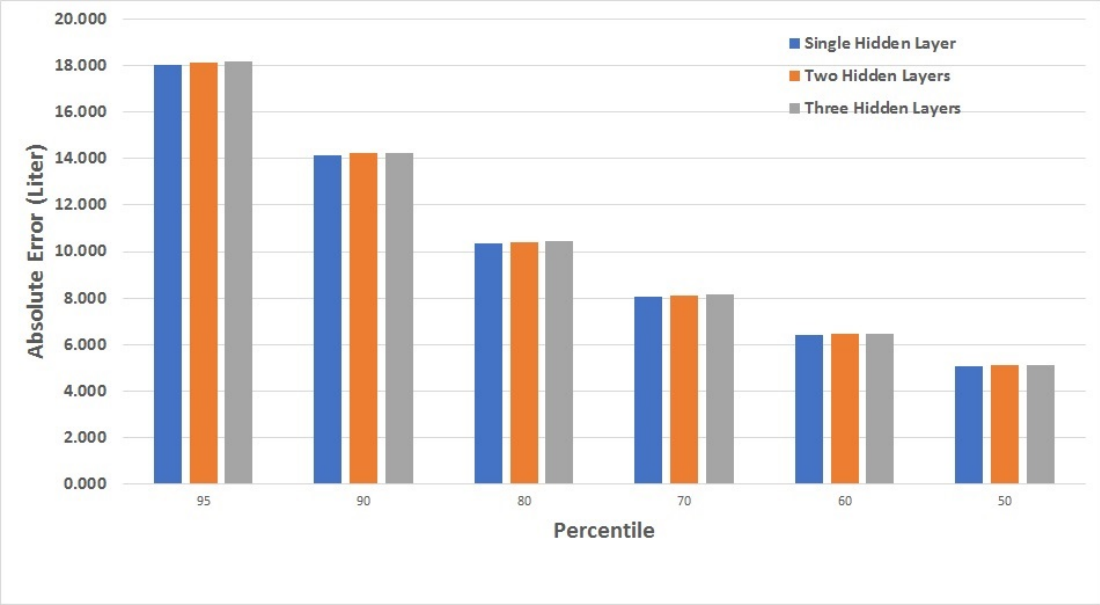


Figure 6.7: One Day Ahead Overall Models Comparisons Weekends AE.

scenarios, under estimated is the most critical cost, since the under estimated predictions could cause insufficient supplies. From all the models, the baseline model is still dominated by all others as it retains the highest percentage of under estimated values. While comparing the one hour ahead and one day ahead models, the one day ahead models retain notable advantages. First, the one day ahead model retains the least percentage of under estimated values for both weekday and weekend scenarios. Moreover, none of the under estimated values are above the threshold. These two advantages show that the one day ahead model outperform the one hour ahead model. Therefore, the one day ahead model is considered the best model over all others.

In addition to comparing overall performance of proposed models, outliers and well predicted instances are investigated. The first analysis is conducted in dissemination area levels. Top 10 outliers and top 10 well predicted dissemination areas are inspected from both an AE and an APE perspective. For outliers, seven out of ten dissemination areas are overlapped across all proposed models. The left three columns in Table 6.7 indicate that all seven dissemination areas retain a low volume of single family clients. Over all 111 experiment dissemination areas, only 17 of them retain single family residents that are less than or equal to 44. This implies that the experiment sample size has a great impact on model predictions. For small population areas, each instance within the group

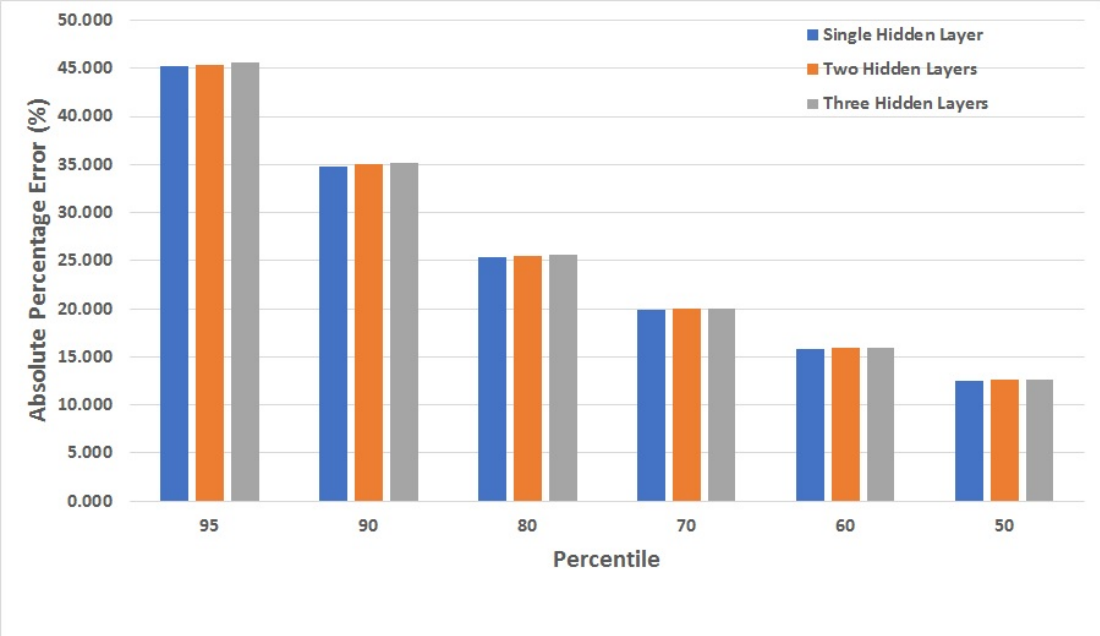


Figure 6.8: One Day Ahead Overall Models Comparisons Weekends APE.

contributes significantly for overall accuracy. If there is an outlier, then predictions of the entire dissemination area will be misled. In addition to outlier dissemination areas, well predicted areas are investigated. 8 out of 10 dissemination areas are in common for all models. Overlapped dissemination area information is demonstrated in the right three columns in Table 6.7. Compared to outlier areas, well predicted areas retain a much larger population. It further confirms the importance of population for water usage predictions. Another impressive observation is that there is a well predicted dissemination area merged by 11 different areas. Although this research is doubtful that merged areas may retain lower performance since factors used to migrate dissemination areas may not well-represent resident water consumption behaviors, this area indicates that once the population reaches a certain amount, varieties of individual impacts can be overcome and models are able to make precise predictions. By contrast, Area3 and Area4 in outlier group constituted by 5 and 3 different areas maintain low accuracy rates. This can be explained by the varieties of different water consumption behaviors, and more importantly by small population. Therefore, another conclusion from this research is that the finer prediction grids are, the more challenging it will to keep accuracy rates at high levels. Suggestions for future works to tackle small population areas are merging areas with the ones close to it and ensuring merged area populations reach to an acceptable amount.

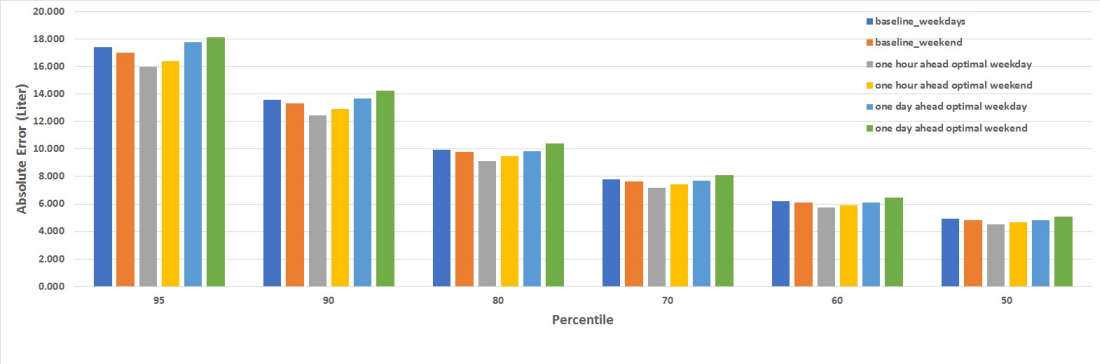


Figure 6.9: AE Comparisons of All Proposed Models.

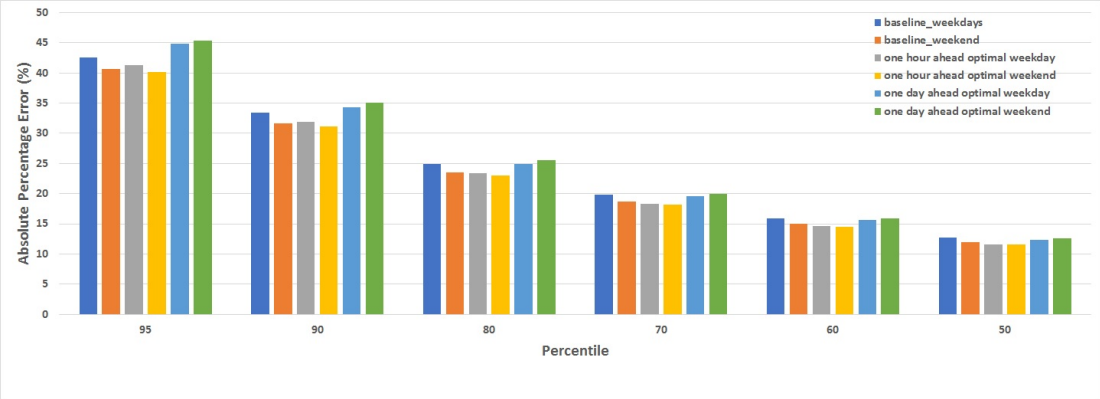


Figure 6.10: APE Comparisons of All Proposed Models.

During seasonal prediction analysis, summer water consumption forecasts are considered more challenging than other seasons because there are more outdoor water consumption activities such as irrigation, car washes and filling swimming pools. Figure 6.11 demonstrates average daily temperatures in degrees over the entire experiment period at Abbotsford. In order to adequately present varieties of outdoor usage over the summer period, this research takes 15 degrees as a threshold and the period between June 22, 2013 and August 30, 2013 as the one to be investigated. As in other experiments, only peak hour predictions are evaluated. In contrast to summer predictions, winter predictions are demonstrated as well. The temperature threshold for winter is 10 degrees and the period considered is from November 6, 2012 to March 12, 2013. In the winter time, major water consumption are dedicated by indoor usage; therefore, they should be more predictable than in the summer months.

Table 6.6: Model performance comparisons by the measurements of over and under estimated predictions.

	overestimated	overestimated above threshold	under estimated	under estimated above threshold	precisely estimated
One hour ahead weekdays	51.7%	51.7%	47.7%	47.7%	0.6%
One hour ahead weekends	50.8%	50.6%	49.2%	49.1%	0.1%
One day ahead weekdays	43.5%	0.0%	43.1%	0.0%	13.4%
One day ahead weekends	43.5%	0.0%	43.2%	0.0%	13.3%
baseline weekdays	47.3%	47.2%	52.7%	52.6%	0.0%
baseline weekends	48.2%	48.1%	51.8%	51.7%	0.0%

Table 6.7: Commonly selected poorly and well predicted dissemination areas.

Outlier Areas			Well Predicted Areas		
Areas	No. SF	No. merged areas	Areas	No. SF	No. merged areas
OutlierArea1	32	1	BestArea1	114	1
OutlierArea2	34	1	BestArea2	117	11
OutlierArea3	35	5	BestArea3	132	1
OutlierArea4	36	3	BestArea4	133	1
OutlierArea5	39	1	BestArea5	141	2
OutlierArea6	42	1	BestArea6	155	1
OutlierArea7	44	1	BestArea7	161	1
			BestArea8	172	2

Figure 6.12 compares average AE among summer, winter and entire experiment periods. There are four groups of data, one for each proposed model. Blue, grey and orange bars represent errors in summer, winter and entire experiment periods respectively. Summer period predictions are significantly worse than the other two, and the performance differences are remarkable. In contrast, winter period predictions are much more accurate than overall and summer predictions. Hence, the study concludes that summer time hourly

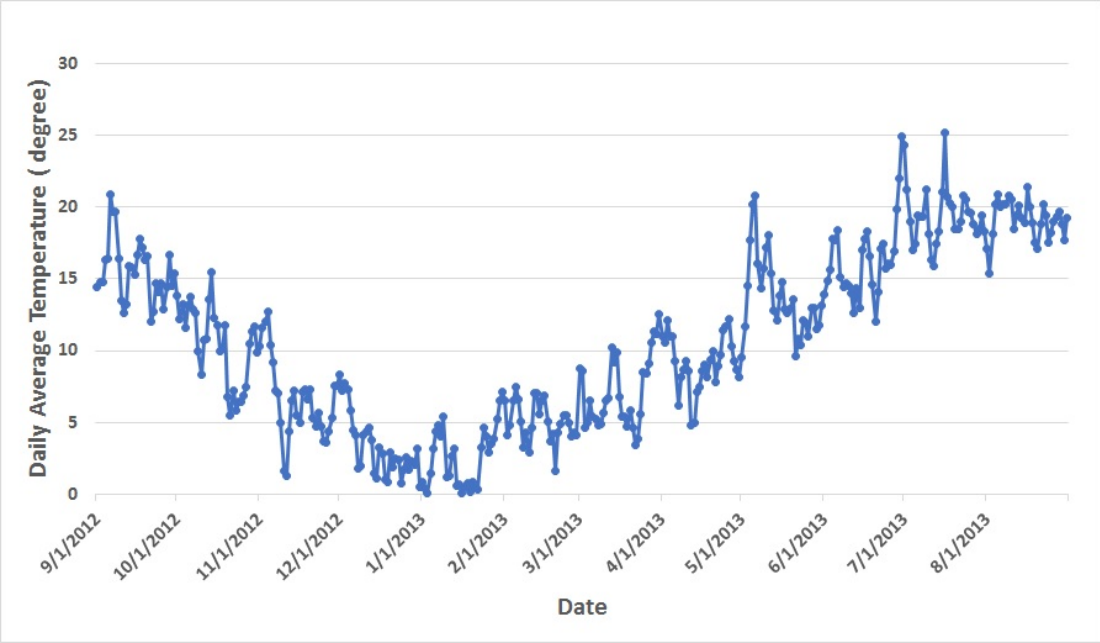


Figure 6.11: Average Daily Temperature.

consumption are more difficult to predict since there are varieties of outdoor consumption behaviors for different families. Moreover, winter period forecasts are more feasible to reach a high precision because indoor water consumption are the major usage and consumption patterns are identical for all families.

6.3 Summary

In this chapter, I demonstrate proposed models for one hour ahead and one day ahead weekdays and weekends. I conclude that feature selection and weekend and weekday separations are pivotal for hourly water predictions. Regarding multi hidden-layer neural networks, I demonstrate that the two hidden-layer model out performs the single and three hidden-layer models. Moreover, there are a few interesting observations and results presented as well. Model performance is highly related to sample sizes of experiments. Merged areas with small populations are shown to lead to low model accuracy; however, the ones with large populations perform well. Lastly, winter and summer forecasts are compared. One conclusion is that winter consumption, which are indoor usage driven, are more likely to be precisely predicted than summer consumption, which are outdoor usage driven.

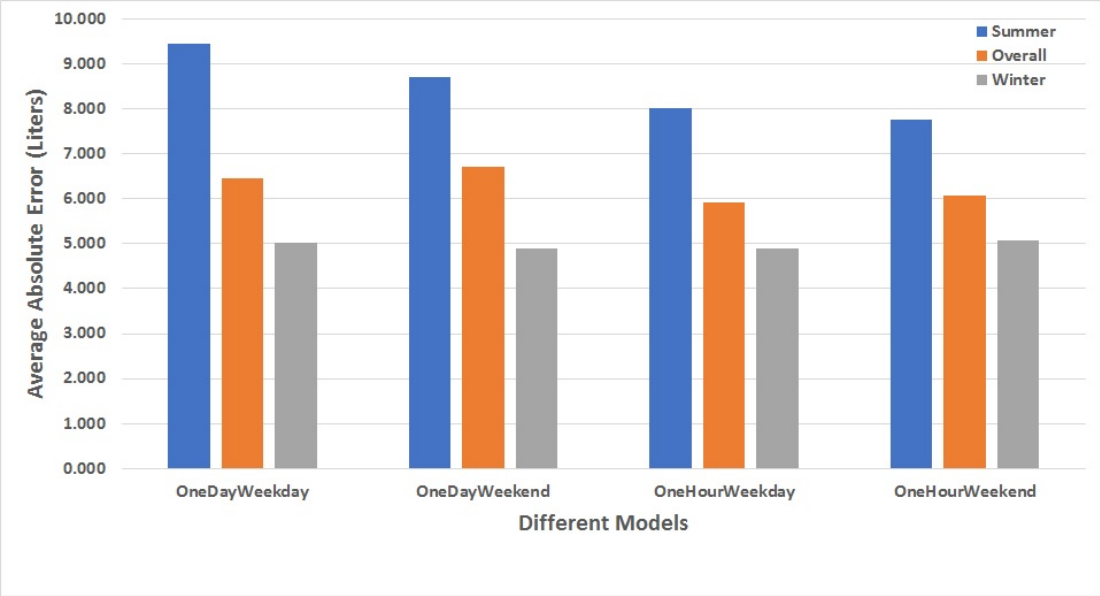


Figure 6.12: AE Comparisons of Summer, Winter and Overall Season.

Chapter 7

Conclusion and Future Work

In this chapter, I summarize the contributions of this thesis and make some suggestions on directions that could further improve model accuracy for future work.

7.1 Conclusion

This research presents artificial neural network models to predict hourly water consumption at the dissemination area level. From the results of small and large population dissemination areas it is clear that the smaller the population is, the more challenging it is to accurately forecast. Therefore, developing precise predictions of water consumption at the dissemination area level is considered as a solid contribution of this thesis. This thesis predicts hourly water consumption in both an hour and a day in advance. One hour ahead predictions are more accurate than one day in advance predictions; however, model performance differences between these two scenarios are at the one cubic liter per resident levels. Due to input data (previous hourly water consumption) availabilities and the accuracy of our models, our one day in advance models could be valuable for utility companies.

In addition, this research demonstrates the benefits and performance improvements that can be obtained by using feature selection and by the separation of weekday and weekend models. In contrast to previous work, this research provides detailed illustrations on determinations of choosing single or separated models containing weekend and weekday residential water usage predictions. Moreover, this thesis not only compares different models from a performance perspective, but also provides each model with a customized feature set by applying feature selection. As Jain and Varshney indicate [33], water consumption

predictions are data dependent. Implementing models with proper feature sets not only improves model accuracy but also reduces model complexities.

Regarding multi hidden-layer models, this research compares single, two and three hidden-layer ANN performance. lthough single hidden layer model can adequately forecast all scenarios in theory, experimental results indicate that models with two hidden-layer models outperform all others from both a performance and a model complexity perspective on predicting water consumption.

Lastly, it is clear that dissemination area population sizes and model performance have a close relationship. The larger the population a dissemination area has, the better model performance will be. For some large population areas, even though they are merged by several dissemination areas, their predictions are still highly accurate. This confirms a contribution of this research, which is predicting usage in a finer grid (dissemination area). Moreover, winter and summer predictions are compared. Winter consumption are normally considered as base usage since all families consume water in similar patterns. In contrast, summer usage may vary as outdoor consumption can be significantly different. The result shows that summer predictions are in a low accuracy rate as compared to overall and winter predictions, which further confirms the challenges.

7.2 Future Work

Although this research constructs models that have a high average accuracy rate, there are some adjustments that could further improve model performance. The first adjustment is collecting weather data in a finer resolution. In this research, daily weather information is collected; however, the water predictions are at an hourly level. If hourly weather information could be collected, it could potentially improve predictions. The second approach is refining the merging algorithm. The merging algorithm in this research is based on the demographic information of each dissemination area, and the population threshold is set to 30. However, dissemination areas with small populations are shown to lead to low accuracy rates compared to large population areas. Hence, leveraging water consumption similarities to merge neighborhood dissemination areas and setting a larger threshold for population sizes may provide better model predictions. The third approach is dividing winter and summer predictions. Summer and winter consumption are presented in significantly different accuracy levels. By following the same approach as weekday and weekend separations, splitting forecasts for winter and summer consumption can lead models to focus on their target period usage patterns which may lead to better model performance. The last approach to improve model accuracy is engaging more water consumption data. In this

research, only one-year water consumption are engaged due to supplied data limitations. However, seasonal trends are normally well predicted by leveraging several years data in previous works. Therefore, engaging previous consecutive years data into experiments may improve prediction accuracy as well.

References

[1] J. Adamowski, H. F. Chan, S. O. Prasher, B. Ozga-Zielinski, and A. Sliusarieva. Comparison of multiple linear and nonlinear regression, autoregressive integrated moving average, artificial neural network, and wavelet artificial neural network methods for urban water demand forecasting in Montreal, Canada. *Water Resource Research*, 48, 2012.

[2] J. F. Adamowski. Peak daily water demand forecast modeling using artificial neural networks. *Journal of Water Resources Planning and Management*, 2:119–128, 2008.

[3] D. E. Agthe and R. B. Billings. Water price influence on apartment complex water use. *Journal of Water Resources Planning and Management*, 128, 2002.

[4] C. K. Aitken, T. A. McMahon, and A. J. Wearing. Residential water use: Predicting and reducing consumption. *Journal of Applied Social Psychology*, 24:136–158, 1994.

[5] M. A. Al-Zahrani and A. Abo-Monasar. Urban residential water demand prediction based on artificial neural networks and time series models. *Water Resources Management*, 29:3651–3662, 2015.

[6] Ethem Alpaydin. *Introduction to Machine Learning*. The MIT Press, 2014.

[7] A. Altunkaynak, M. Özger, and M. Cakmakci. Water consumption prediction of Istanbul city by using fuzzy logic approach. *Journal of Water and Climate Change*, 19:641–654, 2005.

[8] S. Alvisi, M. Franchini, and A. Marinelli. A short-term, pattern-based model for water-demand forecasting. *Journal of Hydroinformaics*, 9:39–50, 2007.

[9] F. Arbués, M. Á. Garcia-Valinas, and R. Martínez-Espi neira. Estimation of residential water demand: A state of the art review. *The Journal of Socio-Economics*, 32:81–102, 2003.

[10] M. S. Babel and V. R. Shinde. Identifying prominent explanatory variables for water demand prediction using artificial neural networks: A case study of Bangkok. *Water Resources Management*, 25:1653–1676, 2011.

[11] M. Bakker, H. van Duist, K. van Schagen, J. Vreeburg, and L. Rietveld. Improving the performance of water demand forecasting models by using weather input. *Procedia Engineering*, 70:93–102, 2014.

[12] C. D. Beal and R. A. Stewart. Identifying residential water end uses underpinning peak day and peak hour demand. *Journal of Water Resources Planning and Management*, 140, 2014.

[13] C. Bennett, R. A. Stewart, and C. D. Beal. ANN-based residential water end use demand forecasting model. *Expert Systems with Applications*, 40:1014–1023, 2013.

[14] E. J. M. Blokker, J. H. G. Vreeburg, and J. C. van Dijk. Simulating residential water demand with a stochastic end-use model. *Journal of Water Resources Planning and Management*, 136, 2010.

[15] J. Bougadis, K. Adamowski, and R. Diduch. Short-term municipal water demand forecasting. *Procedia Engineering*, 19:137–148, 2005.

[16] T. C. Britton, R. A. Stewart, and K. R. O'Halloran. Smart metering: enabler for rapid and effective post meter leakage identification and water loss management. *Journal of Cleaner Production*, 54:166–176, 2013.

[17] H. Chang, G. Hossein Parandvash, and V. Shandas. Spatial variations of single family residential water consumption in Portland, Oregon. *Urban Geography*, 31:953–972, 2010.

[18] W. A. Clark and J. C. Finley. Determinants of water conservation intention in Blagoevgrad, Bulgaria. *Society and Natural Resources*, 20:613–627, 2007.

[19] A. Cominola, M. Giuliani, D. Piga, A. Castelletti, and A. E. Rizzoli. Benefits and challenges of using smart meters for advancing residential water demand modeling and management: A review. *Environmental Modelling and Software*, 72:198–214, 2015.

[20] S. A. Eslamian, S. S. Li, and F. Haghighat. A new multiple regression model for predictions of urban water use. *Sustainable Cities and Society*, 27:419–429, 2006.

[21] M. Firat, M. E. Turan, and M. A. Yurdusev. Comparative analysis of neural network techniques for predicting water consumption time series. *Journal of Hydrology*, 384:46–51, 2010.

[22] C. Fox, B. S. McIntosh, and P. Jeffrey. Classifying households for water demand forecasting using physical property characteristics. *Land Use Policy*, 26:558–568, 2009.

[23] F. Gagliardi, S. Alvisi, M. Franchini, and M. Guidorzi. A comparison between pattern-based and neural network short-term water demand forecasting models. *Water Science and Technology: Water Supply*, 18:1426–1435, 2017.

[24] R. Gargano, C. Tricarico, F. Granata, S. Santopietro, and G. de Marinis. Probabilistic models for the peak residential water demand. *Water*, 9, 2017.

[25] M. Ghiassi, D. K. Zimbra, and H. Saidane. Urban water demand forecasting with a dynamic artificial neural network model. *Journal of Water Resources Planning and Management*, 134:138–146, 2008.

[26] I. Guyon and A. Elisseeff. An introduction to variable and feature selection. *Journal of Machine Learning Research*, 3:1157–1182, 2003.

[27] J. Han, J. Pei, and M. Kamber. *Data Mining Concepts and Techniques*. Morgan Kaufmann, 2011.

[28] R. Hecht-Nielsen. Theory of the backpropagation neural network. *Neural Networks for Perception*, pages 65–93, 1992.

[29] M. Herrera, L. Torgo, J. Izquierdo, and R. Pérez-Garcia. Predictive models for forecasting hourly urban water demand. *Journal of Hydrology*, 387:141–150, 2010.

[30] L. House-Peters, B. Pratt, and H. Chang. Effects of urban spatial structure, sociodemographics, and climate on residential water consumption in Hillsboro, Oregon. *Journal of the American Water Resources Association*, 46:461–472, 2010.

[31] L. A. House-Peters and H. Chang. Urban water demand modeling: Review of concepts, methods, and organizing principles. *Water Resources Research*, 47, 2011.

[32] T. C. Hughes. Peak period design standards for small western U.S. water supply systems. *Journal of the American Water Resources Association*, 16:661–667, 1980.

[33] A. Jain and A. K. Varshney. Short-term water demand forecast modelling at IIT Kanpur using artificial neural networks. *Water Resources Management*, 15:299–321, 2001.

[34] 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*, 46:1130–1137, 2007.

[35] R. C. Balling Jr., P. Gober, and N. Jones. Sensitivity of residential water consumption to variations in climate: An intraurban analysis of Phoenix, Arizona. *Water Resource Research*, 44, 2008.

[36] D. S. Kenney, C. Goemans, R. Klein, J. Lowrey, and K. Reidy. Residential water demand management: Lessons from Aurora, Colorado. *Journal of the American Water Resources Association*, 44:192–207, 2008.

[37] P. Kim. *MATLAB Deep Learning: With Machine Learning, Neural Networks and Artificial Intelligence*. Apress, 2017.

[38] M. Kumar, N. S. Raghuwanshi, R. Singh, and W. W. Wallender. Estimating evapotranspiration using artificial neural networks. *Journal of Irrigation and Drainage Engineering*, 128:224–233, 2002.

[39] K. Y. Lee, Y. T. Cha, and J. H. Park. Short-term load forecasting using an artificial neural network. *IEEE Transactions on Power Systems*, 7:124–132, 1992.

[40] R. J. Lewis. An introduction to classification and regression tree (CART) analysis. In *Annual Meeting of the Society for Academic Emergency Medicine*, 2000.

[41] A. Liu, D. Giurco, and P. Mukheibir. Urban water conservation through customised water and end-use information. *Journal of Cleaner Production*, 112:3164–3175, 2016.

[42] J. Liu, H. G. Savenije, and J. Xu. Forecast of water demand in Weinan City in China using WDF-ANN model. *Physics and Chemistry of the Earth*, 28:219–224, 2003.

[43] I. Maqsood, M. R. Khan, and A. Abraham. An ensemble of neural networks for weather forecasting. *Neural Computing and Applications*, 13:112–122, 2004.

[44] R. Martínez-Espińeira. Residential water demand in the northwest of Spain. *Environmental and Resource Economics*, 21:161–187, 2002.

[45] M. A. Mohandes, S. Rehman, and T. O. Halawani. A neural networks approach for wind speed prediction. *Renewable Energy*, 13:345–354, 1998.

[46] G. Panchal, A. Ganatra, Y. P. Kosta, and D. Panchal. Behaviour analysis of multilayer perceptrons with multiple hidden neurons and hidden layers. *International Journal of Computer Theory and Engineering*, 3:332–237, 2011.

[47] Valerie Platsko. Smart-meter enabled estimation and prediction of outdoor residential water consumption, 2018. Master's thesis, University of Waterloo.

[48] A. S. Polebitski and R. N. Palmer. Seasonal residential water demand forecasting for census tracts. *Journal of Water Resources Planning and Management*, 136:27–36, 2010.

[49] W. Richert and L. P. Coelho. *Building Machine Learning Systems with Python*. Packt Publishing, 2015.

[50] J. Schleich and T. Hillenbrand. Determinants of residential water demand in Germany. *Ecological Economics*, 68:1756–1769, 2009.

[51] V. Shandas, M. Rao, and M. McSharry McGrath. The implications of climate change on residential water use: a micro-scale analysis of Portland (OR), USA. *Journal of Water and Climate Change*, 3:225–238, 2012.

[52] I. Tasadduq, S. Rehman, and K. Bubshait. Application of neural networks for the prediction of hourly mean surface temperatures in Saudi Arabia. *Renewable Energy*, 25:545–554, 2002.

[53] 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*, 119:1419–1428, 2015.

[54] R. M. Willis, R. A. Stewart, and K. Panuwatwanich. Quantifying the influence of environmental and water conservation attitudes on household end use water consumption. *Journal of Environmental Management*, 92:1996–2009, 2011.

[55] I. H. Witten and E. Frank. *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann, 2016.

[56] L. Yu and H. Liu. Feature selection for high-dimensional data: A fast correlation-based filter solution. In *International Conference on Machine Learning*, pages 856–863, 2003.

[57] S. Zhang, C. Zhang, and Q. Yang. Data preparation for data mining. *Applied Artificial Intelligence*, 17:375–381, 2003.

[58] S. L. Zhou, T. A. McMahon, A. Walton, and J. Lewis. Forecasting operational demand for an urban water supply zone. *Journal of Hydrology*, 259:189–202, 2002.

APPENDICES

Appendix A

Table 1: Additional features.

Property	Area	average lot size
	AvgBeds	average bedrooms per household
	AvgFullBath	average full bathrooms per family
	AvgHouseArea	average inside house area per family
	AvgPartBath	average partial bedrooms per household
	AvgStories	average stories per family
	AvgYearBuilt	average building year of houses
	BDR_Occupied_1	percentage of 1 bedroom families in the area
	BDR_Occupied_2	percentage of 2 bedrooms families in the area
	BDR_Occupied_3	percentage of 3 bedrooms families in the area
	BDR_Occupied_4	percentage of 4 bedrooms families in the area
Date	IS_HOLIDAY	whether it is a holiday
	WEEK_DAY	day of the week, value from 1 to 7
Weather	barometer	daily barometer value
	PreviousLastRainFall	number of days since last rainfall before yesterday
	PreviousRainfall	rainfall amount yesterday
	rainfall	rainfall amount today
	SinceLastRainFall	number of days since last rainfall before today
	temperature	max temperature today
	windspeed	windspeed today
Demographic	EDU_Level1	percentage of people with post secondary degree
	EDU_Level2	percentage of people with high school degree
	EDU_Level3	percentage of people not in previous two levels
	Employ_E	employment rate in the area
	Employ_U	unemployment rate in the area
	Tax_B	percentage of people with tax below average
	Tax_T	percentage of people with tax above average
	Tax_Per_Person	median tax per person