UNIVERSITÉ DE MONTRÉAL

UNSUPERVISED LEARNING BASED ON MARKOV CHAIN MODELING OF HOT WATER DEMAND PROCESSES

SHU FAN
DÉPARTEMENT DE GÉNIE ÉLECTRIQUE
ÉCOLE POLYTECHNIQUE DE MONTRÉAL

MÉMOIRE PRÉSENTÉ EN VUE DE L'OBTENTION
DU DIPLÔME DE MAÎTRISE ÈS SCIENCES APPLIQUÉES
(GÉNIE ÉLECTRIQUE)
JUIN 2017

© Shu Fan, 2017.

UNIVERSITÉ DE MONTRÉAL

ÉCOLE POLYTECHNIQUE DE MONTRÉAL

Ce mémoire intitulé :

UNSUPERVISED LEARNING BASED ON MARKOV CHAIN MODELING OF HOT WATER DEMAND PROCESSES

présenté par: FAN Shu
en vue de l'obtention du diplôme de: Maîtrise ès sciences appliquées
a été dûment accepté par le jury d'examen constitué de:

M. SIROIS Frédéric, Ph. D., président
M. MALHAMÉ Roland P., Ph. D., membre et directeur de recherche
M. PARTOVI NIA Vahid, Doctorat, membre et codirecteur de recherche
M. LABIB Richard, Ph. D., membre

DEDICATION

À ma famille.

ACKNOWLEDGEMENTS

I sincerely thank my supervisor Roland Malhamé and my co-supervisor Vahid Partovi Nia for their guidance and encouragement in carrying out this work. I also wish to express my gratitude to other team members of smartDesc project who rendered their help.

I am thankful to the LTE laboratory of IREQ for collecting and providing the data used in this thesis.

RÉSUMÉ

L'ensemble des questions analysées dans ce mémoire dérive d'un important projet de recherche multidisciplinaire appelé smartDESC et réalisé à l'École Polytechnique de Montréal entre les années 2012 et 2016. L'objectif général du projet smartDESC était d'utiliser le stockage associé à certains types de charges d'électricité, et naturellement présent de manière distribuée chez des consommateurs, en vue d'aider à compenser les déséquilibres temporaires entre génération et demande de puissance électrique. Ces derniers sont appelés à devenir de plus en plus fréquents avec la fraction d'énergies renouvelables de type intermittent (énergies solaire et éolienne) dans le mélange de sources d'énergie des réseaux électriques modernes où l'écologie occupe une place de plus en plus importante. Au sein de cet effort général, les chauffe-eau électriques constituent un type de charges d'intérêt particulier vu leur ubiquité et la capacité globale de stockage d'énergie significative à laquelle ils sont associés.

Partant d'un ensemble de mesures rendues anonymes de volumes d'extraction d'eau chaude aux 5 minutes, sur une période de plusieurs mois, et fourni par le laboratoire LTE de l'Institut de recherche d'Hydro-Québec, le but de notre recherche était de développer des algorithmes permettant de regrouper des clients individuels en classes de consommation relativement homogènes et dépendantes à la fois du temps de la journée et du jour de la semaine, dans un objectif subséquent de commande coordonnée. Ce faisant, nous devions faire face à trois défis: (i) automatiser la partition des données en segments temporels de durée suffisante pour être statistiquement significatifs, et durant lesquels les statistiques d'extraction d'eau puissent être considérées comme relativement stationnaires; (ii) À l'intérieur de chaque segment temporel, développer des algorithmes d'estimation de paramètres de modèles de chaînes de Markov à deux états (On et Off) d'extraction d'eau avec un paramètre constant par morceaux de taux moyen d'extraction d'eau dans l'état On; (iii) À la lumière des résultats en (ii), développer des algorithmes de classification des usagers en groupes de consommation relativement proches en termes de propriétés statistiques de consommation, selon l'heure de la journée et le jour de la semaine.

Dans ce mémoire, des outils de la théorie de l'apprentissage machine, de statistiques, et de la théorie des processus stochastiques sont proposés pour répondre aux trois défis en question.

ABSTRACT

The set of problems tackled in this master thesis is an offshoot of a large multidisciplinary research project called smartDESC or smart Distribution Energy Storage Controller, which was carried out at École Polytechnique de Montréal between 2012 and 2016. The general thrust of the smartDESC project was the coordinated use of storage associated with electric loads at customer sites; the objective of this coordination was to smooth out the uncontrolled generation variability brought about by ecologically friendly, yet intermittent, energy sources such as wind and solar. In that global effort, one particular class of loads of interest because of their ubiquity, and their significant overall energy storage capacity, is that of electric water heaters.

We start with a data set consisting of anonymized measurements of hot water extraction volumes in 5 minute samples, over a period of several months, for 73 Quebec households. This data is provided by the LTE laboratory of Institut de recherche d'Hydro-Québec. The goal of the research was to develop approaches to cluster individual users into time of the day and day of the week. We intend to cluster users to relatively homogeneous classes from the point of view of timing and volume of water extraction statistics. Other part of smartDESC is to use these homogeneous clusters to implement coordinated control. In doing so three challenges were to be met: (i) to automate the partition of time of the day into segments of sufficient duration for statistical significance, but relatively stationary hot water extraction statistics; (ii) within each one of the time segments considered, to develop for each user estimation algorithms for two-state (On-Off) Markov chain stochastic models of water extraction with a piecewise constant rate of extraction when On, and validate the results; (iii) In light of the results in (ii), to develop clustering approaches to group users into time of the day and day of the week time intervals where they display relative statistical homogeneity as consumers.

In the master thesis, tools from machine learning, statistics and the theory of stochastic processes are used to propose solutions to each of the above three challenges.

TABLE OF CONTENTS

DÉDICACE iii

ACKNOWLEDGEMENTS iv

RÉSUMÉ v

ABSTRACT vi

TABLE OF CONTENTS vii

LIST OF TABLES ix

LIST OF FIGURES x

LIST OF NOTATIONS xi

CHAPTER 1 INTRODUCTION 1
- 1.1 Data structure 2
- 1.2 Obstacles 2
- 1.3 General objectives 3
- 1.4 Thesis objectives 3
- 1.5 Thesis structure 4

CHAPTER 2 LITERATURE REVIEW 5
- 2.1 Overview of residential energy consumption modeling approaches 5
  - 2.1.1 Top-down approach 5
  - 2.1.2 Bottom-up approach 7
- 2.2 Markov chain and electric water heating load model 10

CHAPTER 3 TIME SEGMENTATION 13
- 3.1 Weekday/Weekend scenarios 13
- 3.2 Time period (slot) division 15
  - 3.2.1 One-dimensional fused lasso approach 15
  - 3.2.2 Training and test sets preparation 20
  - 3.2.3 Regression model 23

CHAPTER 4 PARAMETER ESTIMATION 31
4.1 Alternating renewal process 32
4.2 Theoretical results 32
4.2.1 Moments expressions 32
4.2.2 Equilibrium autocovariance functions 33
4.3 Empirical estimate 35
4.4 Estimation formulae 36
4.5 Direct derivation of $\mathbb{E}_s^{(\mathrm{eq})}[w(r)w(r+t)]$ for exponential case 36
4.5.1 Formula derivation of equilibrium autocovariance function 38
4.5.2 Comparison with general alternating renewal process formula 39
4.6 Simulation 40
4.6.1 Simulation validation of method 40
4.6.2 Single client simulation 42
4.6.3 Error by slots for all clients 43

CHAPTER 5 CLUSTERING RESULTS 45
5.1 Clustering method 45
5.1.1 Mixture model 47
5.1.2 Expectation-maximization (EM) algorithm 47
5.1.3 **mclust** and mixture modeling 48
5.2 Clustering results 49
5.2.1 Clustering on λ_0 and λ_1 49
5.2.2 Clustering c 51

CHAPTER 6 CONCLUSION 56
6.1 Summary 56
6.2 Limitations and future prospects 57

REFERENCES 59

LIST OF TABLES

Table 2.1 Perceived positive and negative attributes of the three major residential energy modeling approaches (Swan and Ugursal, 2009). 10
Table 3.1 Table of weekday test set. 22
Table 3.2 Summary of the linear regression model $\lambda = \alpha_0 + \alpha_1\sigma$. 25
Table 3.3 Time segmentation for homogeneous water demand statistics. 26
Table 4.1 Table of functions. 37
Table 5.3 Clustering results of weekday scenario. 53
Table 5.4 Clustering results of weekend scenario (part 1). 54
Table 5.5 Clustering results of weekend scenario (part 2). 55

LIST OF FIGURES

Figure 2.1 Top-down and bottom-up modeling techniques for estimating the regional or national residential energy consumption (Swan and Ugursal, 2009). 6
Figure 2.2 Transition graph of a two-state Markov chain. 11
Figure 3.1 Average hot water demand evolution in different days of week. 13
Figure 3.2 Methodology diagram part 1. 14
Figure 3.3 Soft-thresholding function. 16
Figure 3.4 Estimation picture for the lasso ($p = 2$). 17
Figure 3.5 Rank sorted by σ^2 (a) weekday (b) weekend. 21
Figure 3.6 Example: $\lambda_5^{\text{obs}} = 9.5$ for client 5. 22
Figure 3.7 Linear regression model. 24
Figure 3.8 Weekday auto-selected λ model. 27
Figure 3.9 Weekday λ^{obs} defined slot cut. 28
Figure 3.10 Weekend auto-selected λ model. 29
Figure 3.11 Weekend λ^{obs} defined slot cut. 30
Figure 4.1 Two state Markov chain. 31
Figure 4.2 Observed (red) and simulation (blue) values of $\hat{\bar{Z}}$, $\hat{\gamma}_0$, $\hat{\gamma}_1$ on weekday. 41
Figure 4.3 Observed (top) and simulation (bottom) data of client 6 on weekday. 43
Figure 4.4 Average observed (black) and simulation (red) data of client 6 on weekday. 44
Figure 4.5 Histogram of slot simulation discrepancies relative to observed data. . 44
Figure 5.1 Methodology diagram part 2. 46
Figure 5.2 Clustering result of slot 1 on weekday. 51
Figure 5.3 Sub clustering results of cluster 1 of slot 1 on weekday. 52

LIST OF NOTATIONS

AIC	Akaike information criterion
BIC	Bayesian information criterion
CDA	Conditional demand analysis
EWH	Electric water heater
GMM	Gaussian mixture model
HDD	Heating degree day
MC	Markov chain
MFG	Mean Field Games
RSS	Residual sum of squares
UEC	Unit energy consumption

Chapter 3: Time segmentation

n	Number of sequence length.
i	As an index, denotes a time index.
j	As an index, denotes a client.
λ	Regularization parameter.
y	Response vector of length n.
β	Coefficient vector.
$\text{sign}(\cdot)$	Sign function.
$\text{soft}(\cdot,\cdot)$	Soft-thresholding function.
Z	Average hot water consumption sequence of length $n = 288$.
Z_i	i-th value of sequence Z.
Z_{ji}	i-th value of client j's sequence Z.
σ_j	Standard deviation of client j's sequence Z.
λ_j^{obs}	Regularization parameter decided by visual inspection for client j.

Chapter 4: Parameter estimation

λ_0	Arrival rate of hot water events.
λ_1	Termination rate of hot water events.
c	Constant hot water extraction rate.
$\mathbb{E}^{(\text{eq})}(\cdot)$	Expectation taken under equilibrium conditions.
$\mathbb{E}_s^{(\text{eq})}(\cdot)$	Laplace transform of $\mathbb{E}^{(\text{eq})}(\cdot)$.
$\mathcal{L}(\cdot)$	Laplace transform of a function.

N	Number of 5-min time units per day, equals to 288.
M	Number of days available in data.
i	As an index, denotes a 5-min time unit.
j	As an index, denotes a day.
Z_i	Total volume consumed over the fixed i-th t=5-min time window, i-th observation of a sequence of length N.
Z	Data matrix of M rows and N columns.
Z_{ji}	Total volume consumed over the fixed i-th t=5-min time window of the j-th day, element at i-th column and j-th row of Z.
$\bar{Z}$	Mean value of a sequence of observations or the matrix Z.
γ_k	Autocovariance of lag k.

Chapter 5: Clustering results

$\lambda_0^{(i)}$	Estimated arrival rate of hot water events for client i.
$\lambda_1^{(i)}$	Estimated termination rate of hot water events.
$c^{(i)}$	Estimated constant hot water extraction rate for client i.
G	Number of Gaussian components.
k	As an index, denotes a Gaussian component.
$\boldsymbol{\mu}_k$	Mean vector of k-th Gaussian component.
$\boldsymbol{\Sigma}_k$	Covariance matrix of k-th Gaussian component.
$N(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$	k-th Gaussian distribution component.

CHAPTER 1 INTRODUCTION

The push for increased renewable energy integration in modern electricity networks is a world wide trend primarily driven by concerns about green gas house emissions resulting from the continuous use of fossil fuels as energy sources. Despite the unsettling changes it may produce in "classical" fossil fuel oriented economies, it is increasingly seen as presenting a significant potential of job creation and economic opportunities. The smartDESC project, funded by the Natural Resources Agency of Canada in 2012, was built around a collaboration between École Polytechnique de Montréal and Hydro-Québec (Sirois et al., 2017). The intermittency of solar and wind energy sources presents significant challenges in the daily operation of electric power systems. The various types of energy storage have consequent usefulness in helping to mitigate the negative effects of that intermittent generation character. The stated goal of the investigators in the project was to develop mathematical and engineering approaches to coordinate the energy storage potential associated with millions of electric loads. Within the power system, these electric loads are naturally present, albeit diverse and geographically dispersed. More specifically the goal was to build an architecture for managing the energy consumption anticipation or deferral potential of large groups of energy storage associated devices such as electric water heaters, air-conditioners, electric space heaters, electric vehicle batteries batteries, swimming pool pumps etc. The envisioned architecture would have a hierarchical structure. The top level (utility or third party aggregator) works with aggregated controllable load dynamic models and non controllable load predictions, as well as solar or wind energy predictions over the forthcoming (few hours) control horizon. It generates (continuously updated) optimal power levels or energy content targets for groups of controlled loads for the next hour. The lower level of the architecture would translate in a decentralized, minimally invasive way, the aggregate goals into microscopic level control actions based on the theory of Mean Field Games, designated as MFG for conciseness. While each load is deciding locally for its contribution, the MFG approach on which it bases its decisions still requires that individuals have stochastic models of their own behaviour, as well as the statistical distribution of model parameters associated with other devices within their control group (i.e. one with sufficient homogeneity to be associated with an overall common power or energy target). The objective of this master's thesis is to illustrate for the particular case of electric water heating loads, and based on the particular experimental data samples on electric water heater water extraction volumes over 5 minute periods for a group of 73 households, as collected and provided by the LTE laboratory of IREQ, how one could:

(i) Use probabilistic modeling and analysis to identify, individual household specific, time inhomogeneous stochastic models of electric water heating load extraction processes,

(ii) Use unsupervised learning approaches to both segment weekdays and weekends into overall piecewise approximately stationary statistics periods, and to cluster different households within the resulting time segments into relatively homogeneous groups within the model parameters space.

1.1 Data structure

The data were collected among 73 households for 10,475 days in total, and only cover the winter season (from November to April) in Quebec. The extracted volume of hot water was recorded every five minutes. Therefore, there are 288 measurements per day (24 h/day$\times$60 min/h$\div$5 min=288 /day). For each day investigated, the household number (client tag), the date and the day of the week were also provided.

1.2 Obstacles

The first challenge lies in the nature of the data which, although it provides information on total 5 minute individual household extracted hot water volumes, fails to specify the type of activity underlying the water extraction. This is unlike more complete data sets reported in the literature (Johnson et al., 2014; Abdallah and Rosenberg, 2012) which besides the volume extracted, also specify the particular activity behind the water extraction (e.g. grooming, laundry, washing dishes, etc.). Different activities are typically associated with different water extraction statistics, and it would then be advantageous to produce a *multi-state* Markov chain model, with distinct states associated with distinct water consuming activity types with a resulting model having better prediction capabilities. Instead, in our raw data, we only had the information about total hot water consumption within successive 5 minute time intervals. So we were forced to distinguish only two states 1, 0 in the class of Markov chain models considered; they respectively indicate the presence or absence of water consumption. Furthermore, it was assumed that, in the active state of the binary Markov chain, the hot water extraction rate is *constant over daily time slots yet to be characterized* (see discussion in the next paragraph), and our task during these time slots, was to estimate the hot water extraction rate, the birth rate of water events and their termination rate.

The next set of challenges lies in the lack of stationarity of water demand process statistics, and the need to cluster consumers into what could be considered as relatively homogeneous consumption classes for subsequent control purposes. For instance, the rate of showers in

the morning or evening tends to be high, while little hot water is consumed in the afternoon. Thus, one has to use the data to identify likely intervals where statistics can be reasonably considered as stationary. A so-called Lasso type based approach (Tibshirani et al., 2011) was devised to achieve that purpose. Three Markov chain parameters (rate of hot water extraction, birth rate of water events, death rate of water events) for individual consumers were then estimated over the intervals of quasi-stationarity. Having chosen to consider that the Markov chain parameters for individuals within a given time slot were realizations of an underlying random parameter vector drawn from a set of possible vector Gaussian distributions, unsupervised learning approaches (Scrucca et al., 2016) were then used to cluster the various individuals into relatively homogeneous consumer groups. Note that the specific identity of consumers in clusters could change from one time interval of quasi-stationarity to the next. This discussion leads us to the formulation of general objectives and specific objectives for our thesis.

1.3 General objectives

The general objective of this research is to establish Markov chain models that would permit a satisfactory anticipation of hot water consumption events over time both for individual power system customers and based on these individual stochastic processes, to anticipate the dynamics of aggregate controlled electric water heating loads.

1.4 Thesis objectives

In order to fully accomplish the general objectives, a set of specific objectives is established:

1. To identify different scenarios to analyse based on the data structure and to determine the appropriate general time segmentation for all clients.

2. To estimate the value of three features (rate of hot water extraction, birth rate and termination rate of water events) for each client and over time intervals where quasi-stationarity holds.

3. To establish the models through the clustering on the three features and find out the most representative ones.

1.5 Thesis structure

Chapter 2 reviews the existing residential energy consumption modeling approaches and lists the previous work on the electric water heating load model. Chapter 3 explains the different scenarios developed in this study and describes the approach for determining the general time segmentation. Chapter 4 presents the methodology of parameter estimation for the three features. Chapter 5 shows how the typical pattern of sub-populations are found through clustering on these features in every time period. The architecture of the methodology in this project is illustrated in Figure 3.2 and Figure 5.1.

CHAPTER 2 LITERATURE REVIEW

In this chapter, we first review the main existing techniques for modeling the residential energy consumption in section 2.1 and then take a look at the Markov chain and the electric water heating load model in section 2.2.

2.1 Overview of residential energy consumption modeling approaches

Renewable energy has attracted much attention over the past decade. During that period, the subject has been extensively explored, and it is still under investigation both in its methodological aspects as well as in energy management applications. A significant number of studies has been devoted to the study of end-use energy on the basis of data collection in the residential sector, for different types of electric loads. In this thesis, our main focus is on the literature related to hot water consumption in electric water heaters (EWH's) given that the emphasis of the smartDESC project of which this research has been a part, has been on learning how to use EWH's as effective batteries. Indeed, EWH's can both help store excess generated power, and contribute to load relief in times of generation deficit by deferring temporarily their own electricity consumption. Although the research background and the consumption categories (water or electricity) may be different, previous studies can be valuable sources of ideas in data management, analysis and data based statistical model building. In this section, we provide an overview of residential end-use energy consumption modeling approaches. Figure 2.1 presents the main existing modeling methods while Table 2.1 lists perceived positive and negative attributes.

2.1.1 Top-down approach

In the top-down approach, the residential sector is regarded as an energy sink, and the individual end-uses details are neglected. Parti and Parti (1980) point out that the number of occupants, electricity price and household income are the three variables which have an effect on the energy consumption related to hot water extraction. The input variables in top-down models usually are the macroeconomic factors (e.g. GDP[1], income, pricing policies), climate-related effects, the number of units of residences (Swan and Ugursal, 2009). These common and widely available variables as well as the historical consumption data are usually used to build one or several equations. Therefore, the residential sector energy needs can be

[1]Gross domestic product.

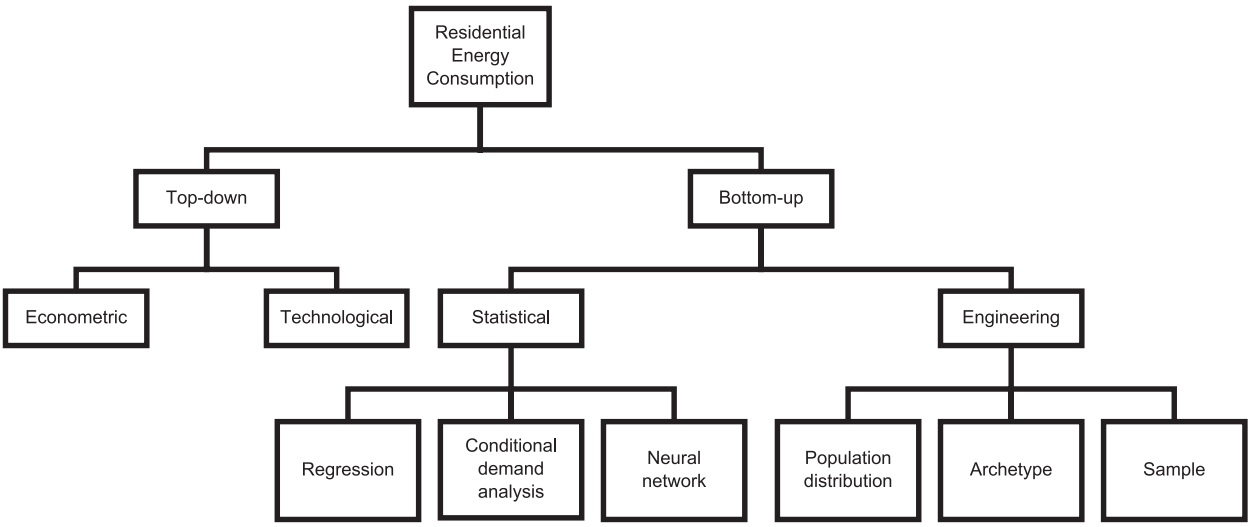


Figure 2.1 Top-down and bottom-up modeling techniques for estimating the regional or national residential energy consumption (Swan and Ugursal, 2009).

estimated approximately. Top-down models fall mainly into two groups: econometric and technological.

Econometric models use mostly the economic factors, such as price and income, as variables. Bentzen and Engsted (2001) proposed and tested Denmark annual energy consumption regression models, which related to only three variables: income, energy price, and heating degree days (HDD[2]).

Technological models use the overall number of residences and the appliance ownership level and average power rating. Zhang (2004) developed the regression equations of unit energy consumption (UEC[3]) using HDD for China, Japan, the United States and Canada respectively and also studied the potential trend of consumption of different types of source of energy.

In top-down approaches, one uses only aggregate energy consumption data, together with a set of relevant macroeconomic indicators, to produce regression models ultimately used for planning purposes. However, their main shortcoming lies in the absence of individual detailed information which prevents a study of the characteristics of typical sub-populations. Moreover, the model is not sensitive to technological advances or sudden housing stock changes, since the regressions are based on the historical consumption information (Swan and Ugursal,

[2]Heating degree day: the number of degrees that a day's average temperature is below 18°C, the temperature below which buildings need to be heated.

[3]Unit energy consumption: the annual amount of energy that is used by the electrical device or appliance.

2009).

2.1.2 Bottom-up approach

In bottom-up approaches, model building starts with an observation of specific end-usages in a set of dwellings, representative of typical sub-populations. Global behaviour is subsequently obtained by aggregating individual dynamics. While more intensive in terms of data collection than top-down approaches, bottom-up approaches can provide load type specific information, as well as aggregate dynamics at substation, regional or national levels depending on the types of applications intended for the models.

Given that the level of details of individual information studied can vary vastly from one dataset to another, the techniques that permit maximum information extraction from the available data also may vary. The so-called *bottom-up statistical* and *bottom-up engineering* are the ones most discussed in the literature. When estimating consumption, the former method is based mainly on historical data, while the latter tends to rely on engineering based knowledge of household appliances, namely their types, power ratings and even the heat transfer characteristics.

Bottom-up engineering model

There are two essential strengths in bottom-up engineering models: (i) ability to reflect technological upgrades and (ii) a much weaker dependence on historical consumption data. Indeed, estimates of energy consumption are obtained by forming the product of expected local consumption as derived from power rating, usage frequency (related to number of members in the household), energy efficiency, heat transfer effect, etc, and the number of houses. When a local technological upgrade occurs (e.g. replacement of heating supply facilities), the engineering model is able to update immediately the new regional and national consumption estimation by simply switching the technological characteristics to the new value. At the "bottom" level, three techniques appear to be predominantly used to capture users variability:

The *distribution technique* uses distributions to represent the appliance ownership among houses and usually picks deterministic averages as characteristics values (Capasso et al., 1994; Kadian et al., 2007).

The *archetype technique* classifies the houses into several categories according to their sizes, year and facilities. Subsequently, the "top" level regional and national consumption is obtained as the sum of the products of number of houses in each category and the energy con-

sumption dynamics of the most typical one, also called archetype, in that category (Huang and Brodrick, 2000).

Finally, the *sample technique* requires the development of a large representative housing energy consumption database to carry out the calculation (Fung et al., 2001).

Although engineering based model building approaches start from the microscopic level using housing stock information, they cannot specify occupants' behaviour pattern in every household. This limitation is precisely due to the fact that engineering-based methods tend to ignore historical consumption data.

Bottom-up statistical model

Statistical model building approaches, which invariably require a large end-use survey sample, are obtained by building individual user models, first including dependency on one or more indicator variables. Such models are essential to retrieve the end-user's behaviour in simulation and estimation. Three model categories are considered, namely polynomial regression models, neural networks and conditional demand models.

Regression models remain the traditional concept. Energy consumption is expressed in terms of a number of observable variables that may have an influence. A variable selection step is sometimes required to simplify and improve the model.

Neural network techniques build a fully computational model. The parameters involved in that nonlinear modeling strategy do not possess any particular physical significance. The neural network consists first of a layer of input variables, then multiple subsequent hidden layers of neurons and a last layer of output energy consumption. A neuron's function is a sum of weighted functions of the neurons in the previous layer. Given a series of observed input/output sequences, the neural network parameters are adjusted by minimization of the output error based on the input of the sequences in the training data set. Aydinalp et al. (2004) developed two neural network based models to estimate the space heating and domestic hot water energy consumptions in the Canadian residential sector. Using the technique presented by Yang et al. (2005) the coefficients and bias in a previously trained network could be continuously updated as new information.

Conditional demand analysis (CDA) techniques rely on the appliance activity record and time-based energy billing data. It quantifies the appliance demand to binary (ON and OFF) or multi-level power rating state. Combined with the energy billing data, it is even possible to study the behaviour pattern for arbitrary appliances and their energy consumption activity. This technique is of particular interest for our research since the electric water heater demand

in this thesis is assumed to be of binary type (either ON or OFF).

Johnson et al. (2014) presented a bottom-up statistical method for modeling household occupant behaviour to simulate residential energy consumption, using a dataset gathered by the U.S. Census Bureau in the American Time Use Survey (ATUS, 2003-2011). The latter defined ten different activities (sleeping, grooming, laundry, food preparation, washing dishes, watching television, using computer, non-power activity, away, away for travelling) as the human behaviour options for the survey and recorded the activity start and stop time based on up to one minute resolution for a total of 124,517 respondents. A Markov chain model was built for each by determining statistically the transition probabilities from one state (or activity) to another at given time t based on the high-resolution data. In the aggregated version, it mentioned that the Markov chain Monte Carlo simulation method required approximately 100 occupants or 40 households for the simulation to produce a reasonably accurate picture of overall residential activity pattern.

One of the most important differences in the nature of our data set and ATUS's or data sets in most other studies is the *availability of activities' label information*. In the majority of researches, the quantity of consumption at time t is collected along with its activity name or related residential load. So it can be modeled with multi-states (distinct activities), and the technological characteristics and performance of activities or appliances (usually using deterministic averages) are also discussed. We note that this thesis is devoted to discussing how to model hot water extraction by means of only a binary Markov chain (extraction present or not) in view of the fact that no information is available as to the reasons of the observed extraction as in the ATUS data set for example.

ATUS data and our research data share the following common points:

- *Both of them employ the concept of quantifying the appliance/activity demand as the states in CDA.*
- *The secondary activities are not taken into account, i.e. both consider that only one activity is taking place on any active time interval.*

Indeed, if the combinations of activities were to be considered, the Markov chain would get more complex, as more situations would have to be taken into consideration, each one being associated with a smaller sample size, and thus with a poorly estimated probability (Johnson et al., 2014). Indeed:

$$N = \sum_{k=0}^{r} \frac{n!}{k!(n-k)!} \tag{2.1}$$

Table 2.1 Perceived positive and negative attributes of the three major residential energy modeling approaches (Swan and Ugursal, 2009).

	Positive attributes	**Negative attributes**
Top-down	• Long term forecasting in the absence of discontinuities • Inclusion of macroeconomic and socioeconomic effects • Simple input information • Encompasses trends	• Reliance on historical consumption information • No explicit representation of end-uses • Coarse analysis
Bottom-up statistical	• Encompasses occupant behaviour • Determination of typical end-use energy contribution • Inclusion of macroeconomic and socioeconomic effects • Uses billing data and simple survey information	• Multicollinearity • Reliance on historical consumption information • Large survey sample to exploit variety
Bottom-up engineering	• Model new technologies • "Ground-up" energy estimation • Determination of end-use qualities based on simulation • Determination of each end-use energy consumption by type, rating, etc.	• Detailed input information • Computationally intensive • Assumption of occupant behaviour and unspecified end-uses • No economic factors

where N the necessary number of states of Markov chain, r the number of activities allowed to take place simultaneously, n the total number of distinct activity types.

At this stage, let us note that the modeling approach adopted in our thesis will draw on both statistical (in particular CDA) and engineering (both archetype and sample) bottom-up modeling approaches, in order to segregate consumers present in our data set into the sub-population classes most relevant for the goals of the smartDESC project, *i.e. the coordination of distributed storage for mitigating generation variability introduced by intermittent renewable energy sources in power systems.*

2.2 Markov chain and electric water heating load model

A Markov chain (MC) is a stochastic process of a particular type. The crucial defining property of Markov chains and which accounts for their wide usage as a tractable stochastic model is *memoryless* i.e., in the case of a discrete time MC, the property that the probability

of moving to different states at time $t+1$ conditional on the state at time t , is independent of all the states that were visited before t (see Lefebvre (2005) for example). As a result, at any time t, one can unambiguously define a matrix of probabilities of transitioning between any two states; it is called the *transition probability matrix*, and it fully characterizes the future probabilistic evolution of a MC, given its most current state.

Thus, assuming S is the state space of a MC of dimension k, the transition matrix P_t is generated in time series of dimension $k \times k$ with the probabilities $p_{i,j} = \Pr(X_{t+1} = j | X_t = i)$ (i.e. the probability of going from state i to state j.). More specifically:

$$P_t = \begin{bmatrix} p_{1,1} & \cdots & p_{1,k} \\ \vdots & \ddots & \vdots \\ p_{k,1} & \cdots & p_{k,k} \end{bmatrix} \qquad \sum_{j=1}^{k} p_{i,j} = 1 \quad \forall i = 1, 2 \ldots, k \tag{2.2}$$

The same information can also be represented by the transition graph. Figure 2.2 is a transition graph of a two-state MC with each state drawn as a circle and the transition probability $p_{i,j}$ drawn as an arrow between states.

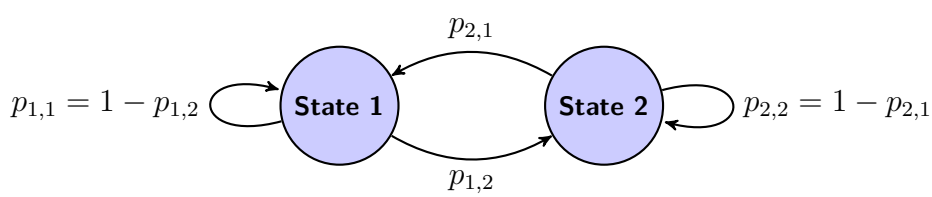


Figure 2.2 Transition graph of a two-state Markov chain.

The elemental electric water heating load model was originally proposed in Chong and Debs (1979). The dynamics in a water heater tank was modeled by using physical thermal characteristics, ambient temperature, inlet/outlet water temperature, state of thermostat control, state of load management control and customer-driven hot water demand. However, the customer-driven hot water withdrawal profile is a non-stationary random process, which is hard to simulate (Kempton, 1988; Alvarez et al., 1992).

Based on the Chong and Debs physically-based elemental stochastic electric water heating load model, Malhamé (1990), and Laurent and Malhamé (1994) have shown how the ideas of statistical mechanics could be used to derive a partial differential equation model of the aggregate behaviour of a large number of identical such loads. Three archetypal households of different physical and consumption parameters were discussed in Laurent et al. (1995). In our research, a two-state MC model is used to simulate the hot water withdrawal profile at the individual device level. The model is described in detail at the beginning of Chapter 4.

Using the ON-OFF two-state MC model, the demand process can also be seen as an alternating 0-1 renewal process with the water extraction rate a constant. Under the assumption that the processes are time homogeneous stationary, El-Férik and Malhamé (1994) developed a methodology to identify the model parameters from the statistical consumption data using the Laplace transform expressions of moments of total occupation time over fixed time windows. Indeed, power company billing is based on measurements of energy consumption data over appropriate fixed length successive time intervals. These fixed-length time intervals were interpreted as the combinations of water heater busy time durations (i.e. with hot water demand present) and that of silent time durations (with no hot water demand). If water extraction, when present, is considered to occur at a constant rate, one can then claim that the total energy extraction is proportional to the total busy time over the fixed interval. The Laplace transform expressions of moments are used in Chapter 4 to establish the equations for estimating the three EWH MC parameters over a given time zone of homogeneous time invariant parameters.

CHAPTER 3 TIME SEGMENTATION

In this chapter, we define different scenarios and cases to study and select the general time segmentation structure, which will be used in the subsequent stage of parameter estimation. Figure 3.2 helps to understand the structure of this chapter, except that the last column is about Chapter 4.

3.1 Weekday/Weekend scenarios

The average plots of hot water consumed, against time in distinct days of a week from Sunday to Saturday, are shown in Figure 3.1. As one can observe from the plots, hot water consumptions on weekdays are significantly different from those on weekends and this result agrees with most people's lifestyles. On weekdays most of the hot water is consumed during two peak periods, in the early morning and in the evening, because consumers are absent or sleeping for the rest of the time. On weekends instead, people tend to get up later, which means starting to use hot water later, and people tend to stay at home longer, which leads to higher water consumption than weekdays. To get a more accurate model, we thus have to

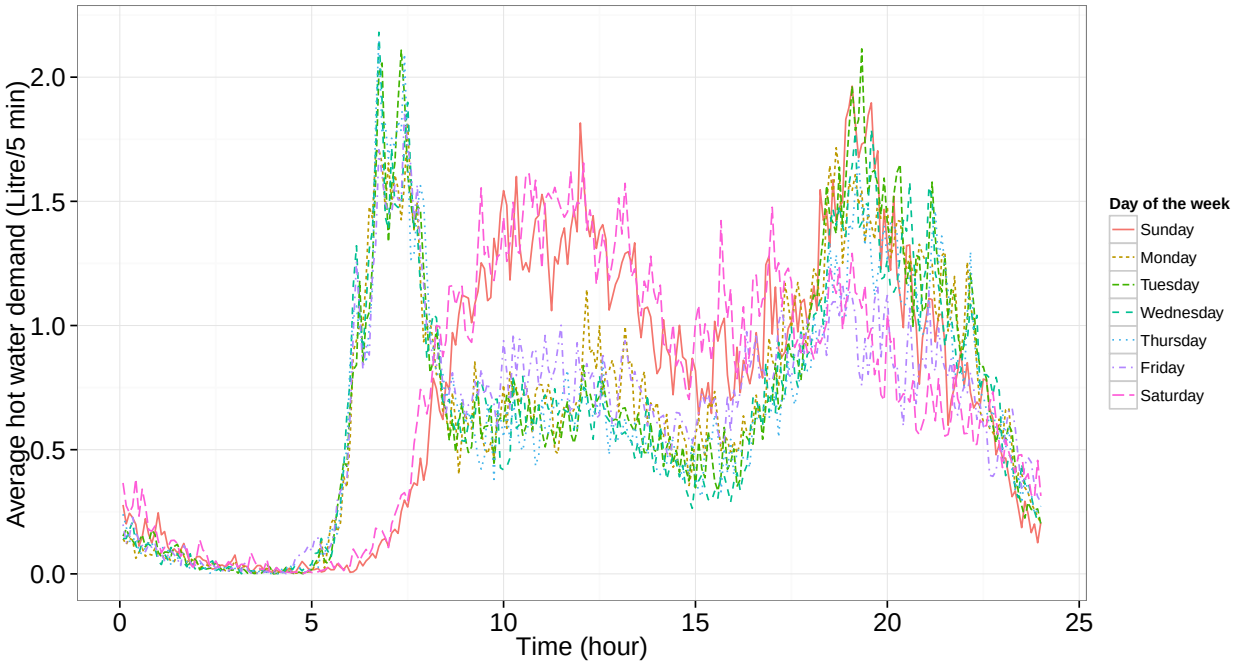


Figure 3.1 Average hot water demand evolution in different days of week.

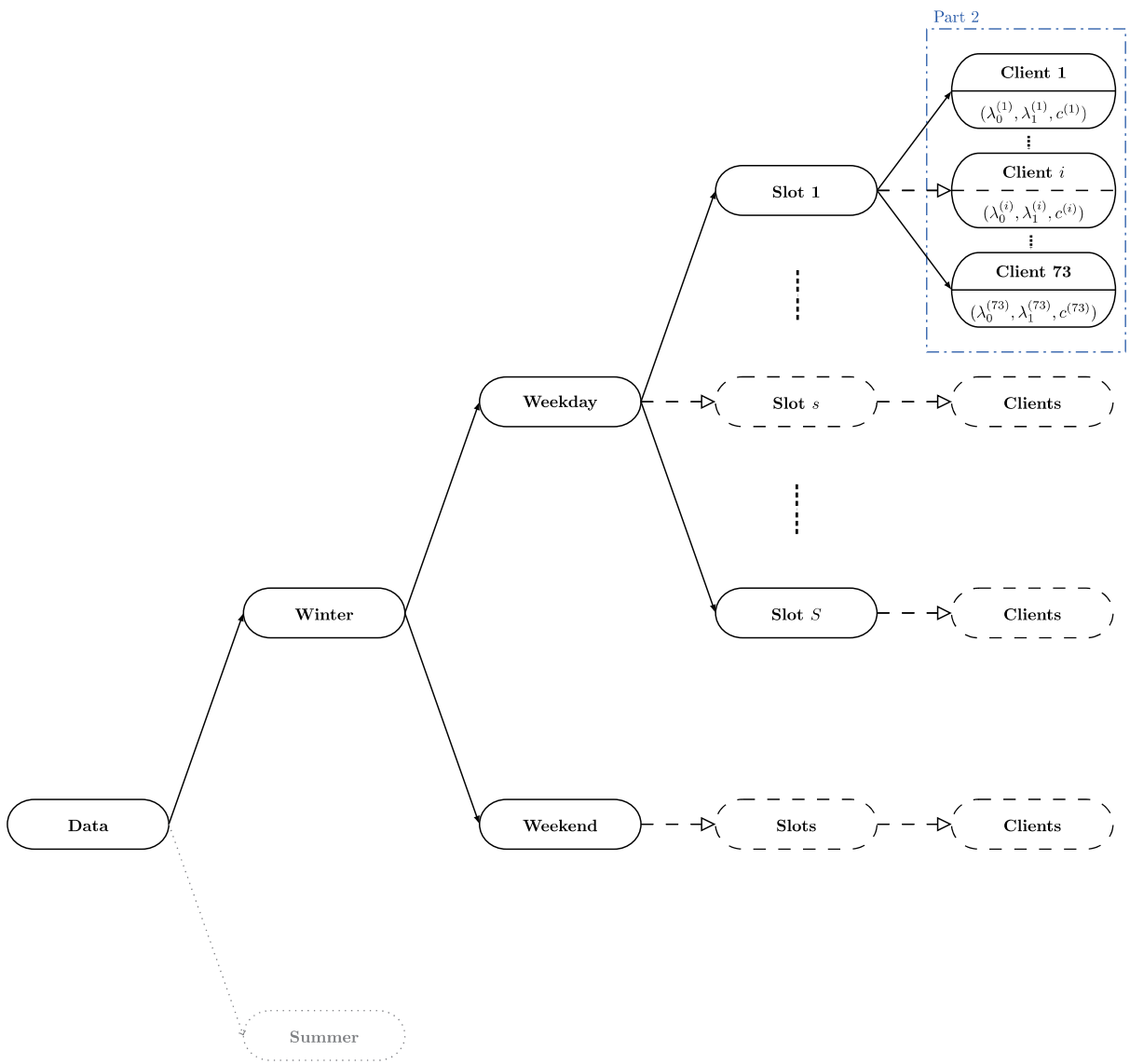


Figure 3.2 Methodology diagram part 1. Starting with the raw dataset, each node represents a subcase defined by a different criterion and the dashed line nodes means there are multiple cases. The directions of the arrows point to the subdivided cases.

deal with the data in weekdays and the data in weekends separately, and these two scenarios are considered in this project. Moreover, the same algorithm and simulation methods are applied to both weekday and weekend scenarios, the only difference being the input data.

The consumption patterns are also expected to vary from one season to the other, particularly for the cities like Montreal with very cold winters, and occasionally very hot summers. Indeed, the temperature is also an important factor affecting people's hot water consumption behaviour. For example, when it gets warmer, people usually tend to take a shorter shower with less hot water, which means there is less energy consumption. There are long summers (from May to October) and long winters (from November to April) in Quebec. The dataset used in this project *only contains data from November to April.* In other words, if we were to continue gathering data during summer time, we would need to further elaborate the model by setting up four scenarios corresponding to weekday/weekend in summer/winter.

3.2 Time period (slot) division

Note that a time inhomogeneous modeling problem is dealt with in this thesis and this property has to be captured in our models. Different distributions are needed to represent the consumption in different periods of time. According to the observed data, most clients maintain similar peak consumption times from day to day, thus leading us to an assumption of 24 hour periodicities related to general human activities. An inhomogeneous but of 24 hour period behavioural model is then broken up into several homogeneous modeling problems.

In order to define the different time slots, we introduce one-dimensional fused lasso (Tibshirani et al., 2011).

3.2.1 One-dimensional fused lasso approach

This section is an introduction about the lasso problems. We start with a minimization problem (3.1) whose result is in the form of soft-thresholding (3.4). It is found that problem (3.1) is somehow equivalent to the common lasso problem (3.5). Then, we specify the one dimensional fused lasso approach and relate it to our case.

Soft-thresholding helps to solve the following optimization problem where both the Euclidean and the absolute value norms are involved

$$\underset{\beta \in \mathbb{R}^n}{\operatorname{argmin}} \|y - \beta\|_2^2 + \lambda \|\beta\|_1 \tag{3.1}$$

with $y = (y_1, y_2, \ldots, y_n) \in \mathbb{R}^n$ is a response vector and $\lambda \geq 0$. The coefficient vector $\hat{\beta} = \hat{\beta}(\lambda)$

is a function of parameter λ. (3.1) can also be written as

$$\operatorname{argmin} \sum_{i=1}^{n} (y_i - \beta_i)^2 + \lambda |\beta_i| \tag{3.2}$$

The sum is minimized whenever one minimizes every component quadratic function of β_i in the sum. The extremum of the function $f(x) = (b - x)^2 + \lambda|x|$ is at $x = b - \lambda \operatorname{sign}(x)/2$. By discussing different conditions of sign of x, values of b, $\lambda/2$ and $f(0)$, the solution of $f(x)$ turns out to be

$$\operatorname{argmin} f(x) = \begin{cases} b + \lambda/2 & \text{if } b < -\lambda/2 \\ 0 & \text{if } b \leq |\lambda|/2 \\ b - \lambda/2 & \text{if } b > \lambda/2 \end{cases} \tag{3.3}$$

It is in a form of soft-thresholding, if we denote b as the variable and $\lambda/2$ as the threshold value. Hence the solution to problem (3.1) for each coordinate $i = 1, \ldots, n$ is

$$\hat{\beta}_i = \operatorname{soft}(y_i, \lambda/2) = \begin{cases} y_i + \lambda/2 & \text{if } y_i < -\lambda/2 \\ 0 & \text{if } y_i \leq |\lambda|/2 \\ y_i - \lambda/2 & \text{if } y_i > \lambda/2 \end{cases} \tag{3.4}$$

where $\operatorname{soft}(\cdot, \cdot)$ is the notation of soft-thresholding and Figure 3.3 shows the function. It is shown that soft-thresholding sets small values $\in [-\lambda/2, \lambda/2]$ to zero while all others are biased.

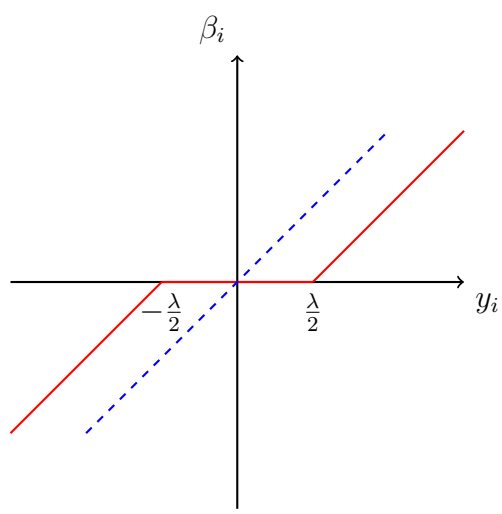


Figure 3.3 Soft-thresholding function. The red line reprsents function $\beta_i = \operatorname{soft}(y_i, \lambda/2)$, while the blue dashed line $\beta_i = y_i$ is the solution of $\operatorname{argmin} \sum_{i=1}^{n} (y_i - \beta_i)^2$.

The lasso problem is commonly written as

$$\hat{\beta} = \underset{\beta \in \mathbb{R}^p}{\operatorname{argmin}} \frac{1}{2} \|y - X\beta\|_2^2 + \lambda \|\beta\|_1 \tag{3.5}$$

where $y \in \mathbb{R}^n$ is a response vector, $X \in \mathbb{R}^{n \times p}$ is a matrix of predictors and $\lambda \geq 0$. The coefficient vector $\hat{\beta} = \hat{\beta}(\lambda)$ is a function of *regularization* parameter λ. When $X = I_{n \times n}$,

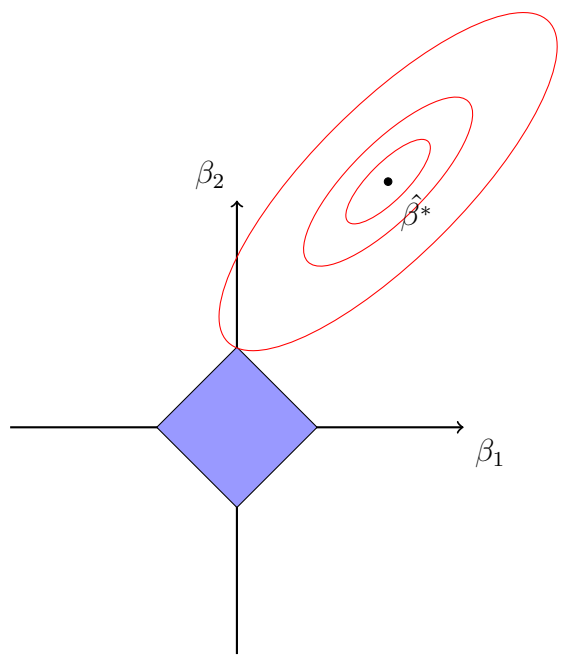


Figure 3.4 Estimation picture for the lasso ($p = 2$). The solid blue area is the ℓ_1 penalty term shown as the constraint region, while the red ellipses are the common residual sum of squares (RSS) contours with $\hat{\beta}^*$ the RSS coefficients. The solution is the first place the RSS contours hit the constraint region.

(3.5) is called signal approximation case and can be rewritten as

$$\begin{aligned} \underset{\beta \in \mathbb{R}^n}{\operatorname{argmin}} \frac{1}{2} \|y - \beta\|_2^2 + \lambda \|\beta\|_1 &= \operatorname{argmin} \frac{1}{2} \sum_{i=1}^{n} (y_i - \beta_i)^2 + \lambda |\beta_i| && (3.6) \\ &= \operatorname{argmin} \sum_{i=1}^{n} (y_i - \beta_i)^2 + 2\lambda |\beta_i| && (3.7) \end{aligned}$$

(3.2) and (3.7) share similar form and the solution of (3.7) is $\operatorname{soft}(y_i, \lambda)$. Figure 3.4 also shows geometrically why the lasso encourages sparse solutions: there are sharp edges and corners for the blue area because of the absolute value in ℓ_1 norm and it is highly likely that the contour will first hit the corner, then some of the coefficients will be set to zero. Hence, the lasso performs shrinkage and subset selection.

Now if the lasso problem (3.5) is generalized to

$$\hat{\beta} = \underset{\beta \in \mathbb{R}^n}{\operatorname{argmin}} \frac{1}{2} \|y - X\beta\|_2^2 + \lambda \|D\beta\|_1 \tag{3.8}$$

where $D \in \mathbb{R}^{m \times p}$ is a penalty matrix. If $X = I$ and D is specified as the $(n-1) \times n$ matrix of first differences given in (3.9), it is called the one-dimensional fused lasso (1d fused lasso) problem.

$$D = \begin{bmatrix} -1 & 1 & 0 & \cdots & 0 & 0 \\ 0 & -1 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & -1 & 1 \end{bmatrix} \tag{3.9}$$

As $X = I$ is full column rank, the solution is

$$\hat{\beta} = y - D^\top \hat{u} \tag{3.10}$$

and

$$\hat{u} = \underset{u \in \mathbb{R}^{n-1}}{\operatorname{argmin}} ||y - D^\top u||_2^2 \quad \text{subject to } ||u||_\infty \leq \lambda \tag{3.11}$$

Furthermore, for each coordinate $i = 1, \ldots, n-1$

$$\hat{u}_i \in \begin{cases} \{-\lambda\} & \text{if } \beta_{i+1} - \beta_i < 0 \\ [-\lambda, +\lambda] & \text{if } \beta_{i+1} - \beta_i = 0 \\ \{+\lambda\} & \text{if } \beta_{i+1} - \beta_i > 0 \end{cases} \tag{3.12}$$

(Arnold and Tibshirani; Tibshirani et al., 2011)

The 1d fused lasso is the common signal approximator case and is used in settings where coordinates in the true model are closely related to their neighbours. Recall the idea of breaking up the inhomogeneous 24 hour period behavioural model into several homogeneous modeling problems. If we take $n = 288$ (see data structure in Section 1.1) and the response vector $y \in \mathbb{R}^n$ as the consumption sequence $Z = \{Z_1, Z_2, \ldots, Z_{288}\}$. Z_i is the volume of hot water consumed during the i-th 5 min of one day ($i = 1, 2, \ldots, 288$). The problem is of 1-dimensional structure and its coordinates are the i-th unit time corresponding to successive

positions on a straight line. The expression of 1d fused lasso becomes

$$\hat{\beta} = \underset{\beta \in \mathbb{R}^n}{\operatorname{argmin}} \frac{1}{2} \sum_{i=1}^{288} (Z_i - \beta_i)^2 + \lambda \sum_{i=1}^{288-1} |\beta_{i+1} - \beta_i| \tag{3.13}$$

The consumption sequence Z is generated from a process whose mean changes at only a smaller number of locations, which implies different human activity time periods. So the 1d fused lasso aims to find a piecewise constant vector β of model coefficients fitting well to consumption sequence Z and the subset selection is the coordinates of changes of β, which indicate the edges of time periods.

Indeed, $\hat{\beta}$ is a function of λ. The ℓ_1 norm penalizes the absolute differences in adjacent coordinates of β. The choice of cost function (3.13) for the segmentation problem, and the role of the lambda coefficient are best understood by considering two limiting cases:

(i) If $\lambda = 0$, (3.13) equals to $\operatorname{argmin} \sum_{i=1}^{288} (Z_i - \beta_i)^2$, so $\hat{\beta}$ is identical with Z. The fitting is perfect but makes no sense, because every data point would define a separate interval.

(ii) If λ approaches infinity, (3.13) is equivalent to $\operatorname{argmin} \sum_{i=1}^{288} (Z_i - \beta_{\text{const}})^2$, where $|\beta_{i+1} - \beta_i| = 0$ and $\hat{\beta}$ becomes a *constant* vector. The best fit is $\beta_{\text{const}} = \sum_{i=1}^{288} Z_i/288$, and all the coefficients become equal to the mean consumption. The 288 degrees of freedom shrink to a single degree of freedom.

So tuning the regularization parameter λ for the 1d fused lasso problem is actually a compromise between accuracy of fit and number of changes between successive values of β.

The crucial point of this method is to choose an appropriate regularization parameter λ. The paragraph "record λ^{obs}" in section 3.2.2 presents how we choose λ for each client. However, contrary to expectation, it is very hard to take this decision directly and work has to be repeated when the sequences change. The cross-validation (Geisser, 1993; Kohavi, 1995; Devijver and Kittler, 1982) is an often used method to automate the choice of $\hat{\lambda}$. However, when the number of clients increases, there are a larger number of sequences to deal with, and more $\hat{\lambda}$ have to be chosen. It would be computationally expensive to use cross-validation every time. So a linear regression model is suggested here rather than the cross-validation. More precisely, in view that larger variance means one needs more effort on trend filtering, and thus λ leading to desirable results has to be higher, we conjecture a linear model relating an "adequate" choice of λ to the standard deviation σ of water consumption over the time slot one has to segment. The linear function parameters will be estimated from a regression analysis based on a preliminary "manual tuning" of λ for a series of examples from some

training set. With σ easily computed from the raw data, this approach can lead to an automation of the choice of $\hat{\lambda}$.

So in the following sections, we present our method to automate the selection of the λ coefficient in the performance criterion. As mentioned above, the strategy is to develop an empirical linear relation between σ and λ and test the resulting performance. First, we separate the data into a training set and a test set. The elements in each of the sets are then separated according to their σ; we choose λ *visually* (manual tuning) for each client (Section 3.2.2). The training set is used to obtain an empirical λ-σ function using linear regression and the test set is subsequently used to evaluate the quality of the time segmentation based on the empirical function obtained. Once the function is defined as acceptable, the linear equation $\lambda = \alpha_1\sigma + \alpha_0$ is used to determine the proper time segmentation for each client (Section 3.2.3).

3.2.2 Training and test sets preparation

Standard deviation

For each client j, $\hat{\sigma}_j$ is defined as

$$\hat{\sigma}_j = \sqrt{\mathbb{V}[Z_j]} \triangleq \sqrt{\frac{1}{288-1}\sum_{j=1}^{288}(Z_{ji} - \bar{Z}_j)^2} \tag{3.14}$$

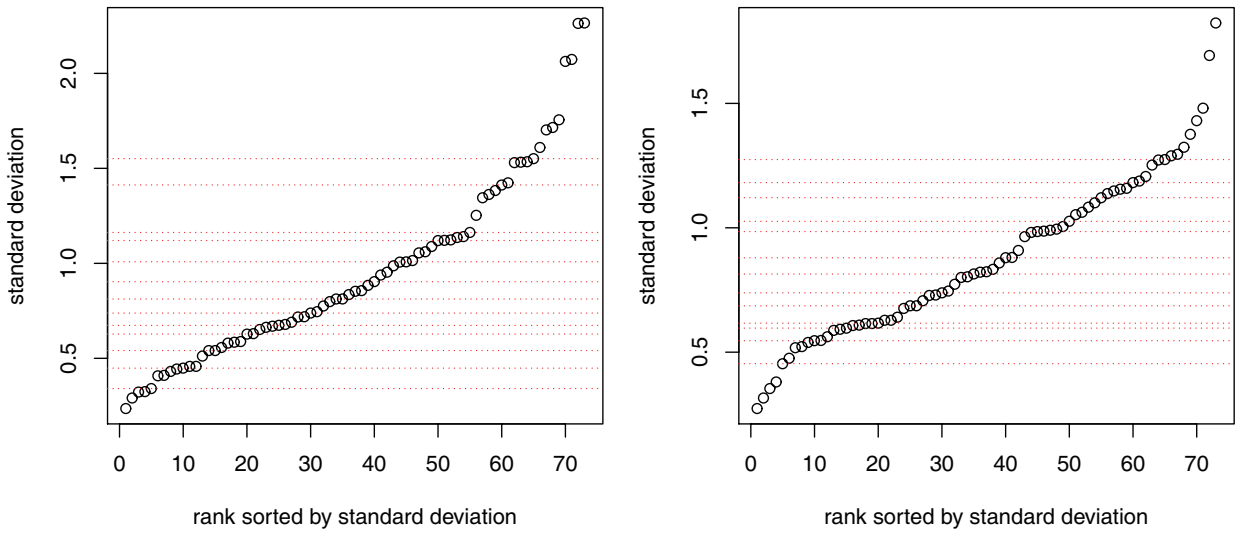
where j is the client tag ($1 \sim 73$). We call i the time index (1∼288). Its maximal value 288 comes from the total number of 5 min unit time per day: 24 h/day $\times$ 60 min/h $\div$ 5 min = 288/day. Z_{ij} represents the average consumption of client j during the i-th 5 min time unit of 24 hours, $Z_j = \{Z_{j1}, Z_{j2}, \ldots, Z_{j288}\}$ and $\bar{Z}_j = \frac{1}{288}\sum_{i=1}^{288} Z_{ji}$.

Figure 3.5 shows 73 clients' standard deviation σ_j of average consumption every 5 min in weekday and weekend scenarios respectively and the decision of forming training set and test set.

Record λ^{obs}

In the fused lasso approach, values of the regularization parameter λ must be chosen for the various customers. This choice is highly dependent on personal discernment. For each client j, several different λ should be tested until we judge the result to be "suitable" and record it as observed value λ_j^{obs}.

Figure 3.6 takes client 5 as an example and $\lambda_5^{\text{obs}} = 9.5$ was chosen to be the observed value



(a) (b)

Figure 3.5 Rank sorted by σ^2 (a) weekday (b) weekend. There are 73 clients' standard deviation σ_j of average consumption every 5 min in weekday and weekend scenarios respectively. y-axis is standard deviation σ_j. x-axis represents their rank according to size among 73 clients. The red lines cross the points that are selected to form a test set with a quantity of 13 in each scenario. As the total number of clients is 73, we choose rank 5, 10, 15, 20, 25, 30, 35, 40, 45, 50, 55, 60, 65. The remaining 60 samples form the training set.

of regularization parameter. There are some characteristics: two peak times present in the morning and evening, low consumption between them and silence during night time.

The black line represents the smoothed signal $\hat{\beta}$ from the 1d fused lasso (3.13). According to (3.10), (3.11) and (3.12), the coefficients are soft-thresholded on difference of adjacent β, so we note this piecewise constant β^{biased}. The time periods are separated by the red vertical lines, where $\beta_{i+1} \neq \beta_i$. The blue dashed line, called β^{unbiased} represents the mean of data whose β^{biased} maintain the same value. It helps to visualize the consumption trend between successive time periods.

The detected positions of changes of β (red) are sometimes consecutive, especially when λ is small. The reason is that there is a gradual increase (decrease) of demand during that period. If we segment a slot containing gradual increase (decrease) demands, it will not be a stationary slot. So λ^{obs} should help decide separate successive regions but be such that the least number of very short slots are detected.

This work is repeated for every client j and we record its λ_j^{obs}. Table 3.1 lists the results obtained in weekday test set.

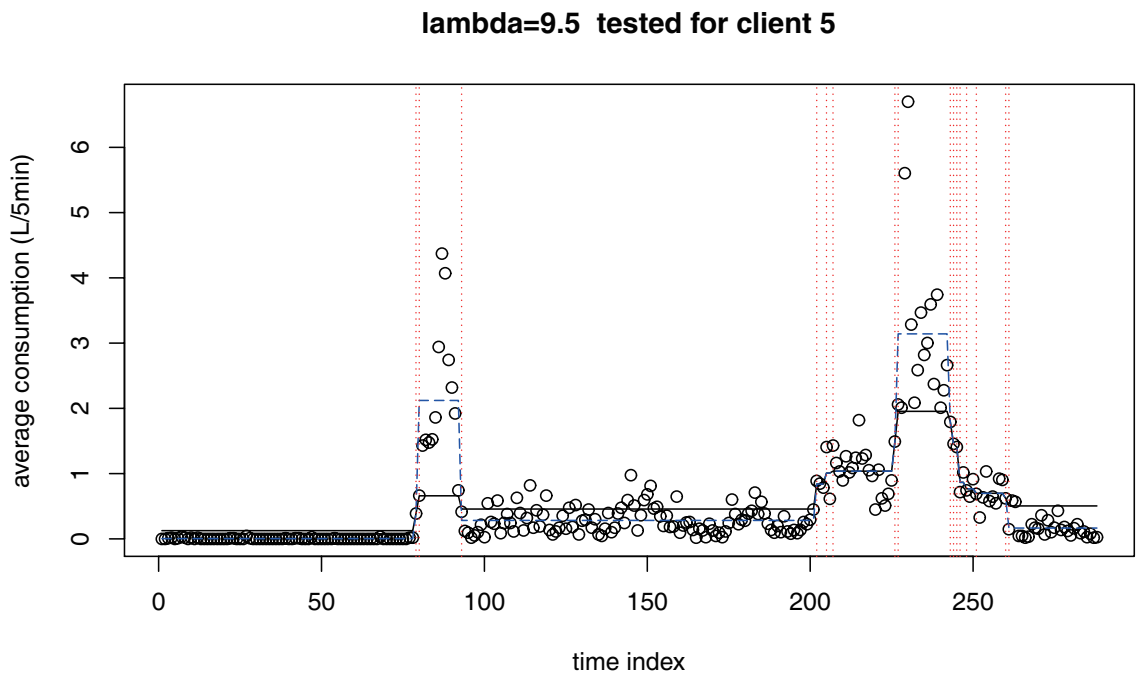


Figure 3.6 Example: $\lambda_5^{\text{obs}} = 9.5$ for client 5. x-axis indicates the sequence number i of unit time 5 min. The black dots represent the average consumption Z_{5i} observed from the data. The black line shows the piecewise constant β^{biased} while the blue dash line is β^{unbiased}. The red dash lines indicate the time period (slot) segmentation.

Table 3.1 Table of weekday test set.

No.	**Client tag** j	**Rank**	**Standard deviation** σ	λ_j^{obs}
1	18	5	0.3	2
2	20	10	0.4	3.1
3	3	15	0.5	3.7
4	73	20	0.6	5.6
5	46	25	0.7	7.8
6	53	30	0.7	7.7
7	22	35	0.8	3.3
8	5	40	0.9	9.5
9	29	45	1.0	11.1
10	14	50	1.1	6.1
11	63	55	1.2	12.4
12	48	60	1.4	14.5
13	34	65	1.6	19.1

3.2.3 Regression model

A linear regression for the choice of $\hat{\lambda}$ is preferable to cross-validation in this case. Because when the number of clients increases, it would be computationally expensive to use cross-validation every time. Moreover, larger variance σ^2 means one needs more effort on trend filtering, and thus λ leading to desirable results has to be higher. We conjecture a linear model relating the choice of $\hat{\lambda}$ to the standard deviation σ of water consumption over the time slot one has to segment. With σ easily computed from the raw data, this approach can lead to an automation of the choice of $\hat{\lambda}$. Also, the linear relation is postulated between λ and the standard deviation σ, not the variance σ^2, because λ and σ have the same scale.

We now study the linear relationship between σ and λ. For that purpose, there is no need to segregate between weekday samples and weekend samples. Therefore, we take the union of weekday and weekend training/test sets as our final training/test set in this linear regression model study.

Model discussion

Figure 3.7 shows "observed" value λ versus sample standard deviation. The linear regression model is in the form of $\lambda = \alpha_1\hat{\sigma} + \alpha_0$ with α_1, α_0 constant. As we use sample standard deviation $\hat{\sigma}$, not standard deviation σ, λ should also be replaced by its estimate $\hat{\lambda}$. Then the following relation is obtained.

$$\hat{\lambda} = 9.4\hat{\sigma} - 1.2 \tag{3.15}$$

Table 3.2 summarizes the results of our example linear regression as well as the R^2 indicator (0.714 a good fit) with a p-value 0.001 (close to zero). The coefficient of determination R^2 is expressed as the ratio of the explained variance to the total variance (Gujarati, 2009). The explained variance is the variance of the model predictions, while the sample variance is the variance of the dependent variable.

$$R^2 = \frac{SS_{\text{reg}}}{SS_{\text{tot}}} = \frac{SS_{\text{reg}}/n}{SS_{\text{tot}}/n} \tag{3.16}$$

where $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$ mean of observed data, $SS_{\text{tot}} = \sum_i (y_i - \bar{y})^2$ total sum of squares and $SS_{\text{reg}} = \sum_i (\hat{y}_i - \bar{y})^2$ regression sum of squares.

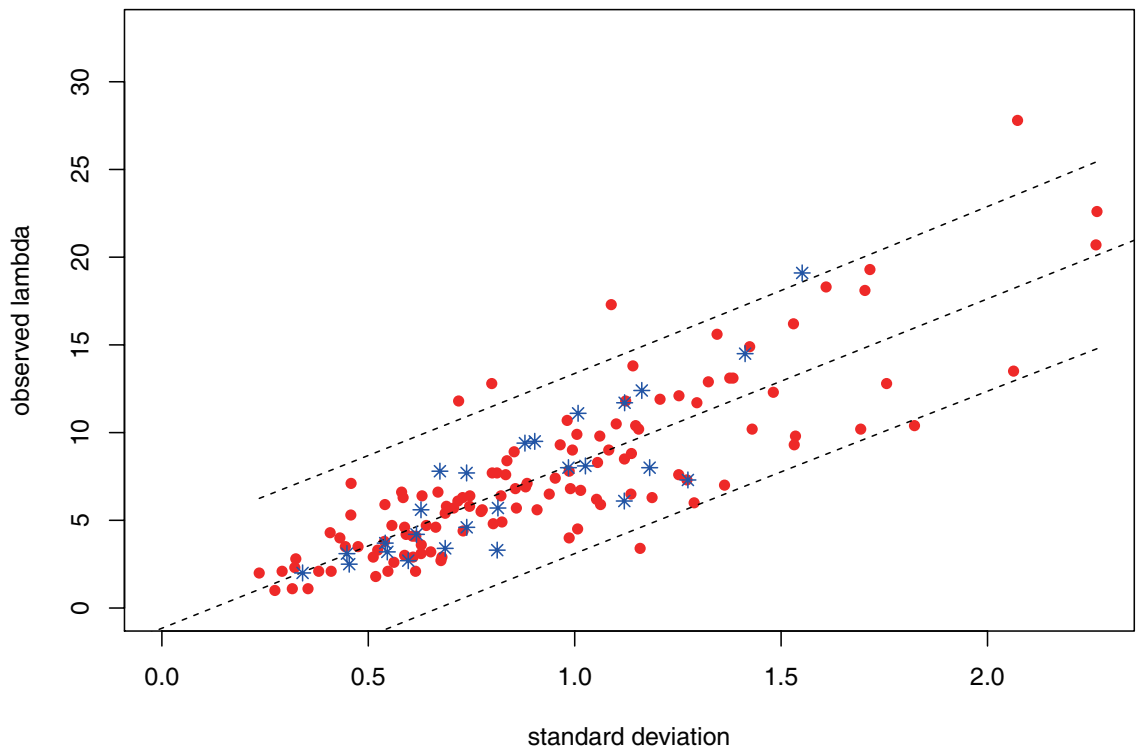


Figure 3.7 Linear regression model. The training set ($60 \times 2 = 120$ points) is plotted in red and the test set ($13 \times 2 = 26$ points) in blue. Note that the blue dots are obtained using independent observations. The three dashed lines are the model obtained $\hat{\lambda} = 9.4\hat{\sigma} - 1.2$, together with the 95% lower and upper confidence bounds. The confidence bounds cover the majority of cases.

Model-based auto-selected λ plotting

This section presents the 73 clients' slot cut visualization. They are

Figure 3.8: Weekday auto-selected λ model

Figure 3.9: Weekday λ^{obs} defined slot cut

Figure 3.10: Weekend auto-selected λ model

Figure 3.11: Weekend λ^{obs} defined slot cut

In the top figures on each page, the black and white colors show different slot cut decided by 1d fused lasso, where their regularization parameters $\hat{\lambda}_j$ are computed from sample standard deviations $sigma_j$ and equation (3.15) for Figure 3.8 and Figure 3.10 or visually decided λ^{obs} for Figure 3.9 and Figure 3.11. Every client has the piece-wise constant vector β^{unbiased}, the edge points of regions are the changing positions from black to white.

Table 3.2 Summary of the linear regression model $\lambda = \alpha_0 + \alpha_1\sigma$.

	Dependent variable:
	λ
Sample standard deviation $\hat{\sigma}$	9.389^{***}
	(0.548)
Constant	-1.154^{**}
	(0.562)
Observations	120
R^2	0.714
Adjusted R^2	0.711
Residual Std. Error	2.581 (df = 118)
F Statistic	293.901^{***} (df = 1; 118)
Note:	*p<0.1; **p<0.05; ***p<0.01

Discussion of results

Our analysis of the results shown in Figure 3.8, Figure 3.9, Figure 3.10, and Figure 3.11 can be summarized in the following two points:

- *The auto-selected λ method offers a satisfactory result.*

In both of the weekday and weekend scenarios, the model based results (Figure 3.8 and Figure 3.10) concur with the results of λ^{obs} (Figure 3.9 and Figure 3.11), most notably in the high similarity between their bar-plots. So our proposed method is verified through this experimental model. Thus, compared with the repeated work of tuning visually λ^{obs} for every sample, if the linear model is well trained, the automatic λ selection method is a much faster and much more convenient way of achieving a good segmentation result, since the standard deviation of data is easy to calculate. If there is a large number of samples to deal with in the future, it is advised to use this method.

- *A general segmentation structure should be selected.*

It is important to remember that modeling effort is aimed at the general objectives of: (i) clustering customers into classes within which consumption patterns are relatively homogeneous; (ii) identifying for each class, the distribution of piecewise constant parameters characterizing the water demand assumed, as we shall see, to evolve according to two-state

Markov chains. In particular, short time intervals (less than 2 hours) are avoided. This is because, in view of the rarity of samples on short intervals, it would be impossible to obtain consistent estimates of the process parameters (estimation is discussed in the next chapter). Also, since we shall see lag 1 autocovariance function empirical estimates are used in the parameter estimation part, sufficiently long time intervals are needed to empirically estimate the value (see equation (4.18) in Section 4.3).

For that purpose, and based on our customer wise time segmentation results, we have to extract common time periods within which a large number of customers are assumed to behave with stationary water consumption statistics. So the boundary points should be set where the majority of clients are changing their consumption pattern. Our analysis of Figure 3.8, Figure 3.9, Figure 3.10, and Figure 3.11 suggest the following time slot sequence of boundary points:

1. weekday time slot boundary points: 72, 96, 156, 216, 264, 288
2. weekend time slot boundary points: 84, 120, 156, 216, 264, 288

This segmentation preserve the intuition of the presence of a silent sleeping time, a morning peak period, a low consumption period, an evening peak period, followed by a quieter early night period.

Table 3.3 Time segmentation for homogeneous water demand statistics.

Slot No.	**Time interval**	
	Weekday	**Weekend**
1	0∼6h	0∼7h
2	6∼8h	7∼10h
3	8∼13h	10∼13h
4	13∼18h	13∼18h
5	18∼22h	18∼22h
6	22∼24h	22∼24h

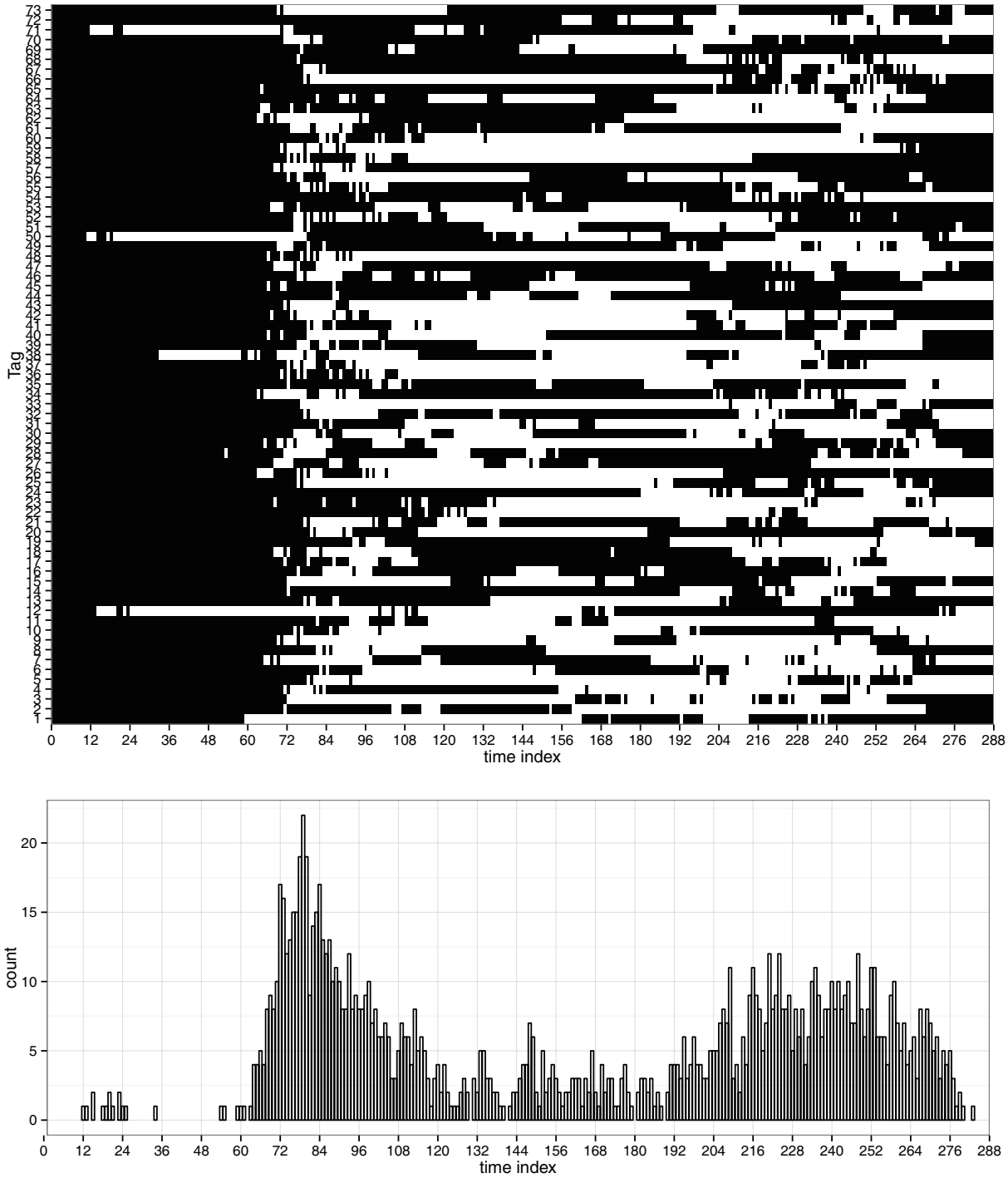


Figure 3.8 Weekday auto-selected λ model. Top panel: x-axis is time sequence number from 1 to 288 (5 min per unit). y-axis is client tag from 1 to 73. For a client y, when the color changes at time index x, it means this client's consumption pattern is changing and we are moving to a new slot. Conversely, when the color remains the same, it means the process is considered to have a relatively stationary property, indicating that parameter estimation theory based on this stochastic assumption is applicable. Bottom panel: The bar-plot indicates how many clients possess a slot cut point at time index x.

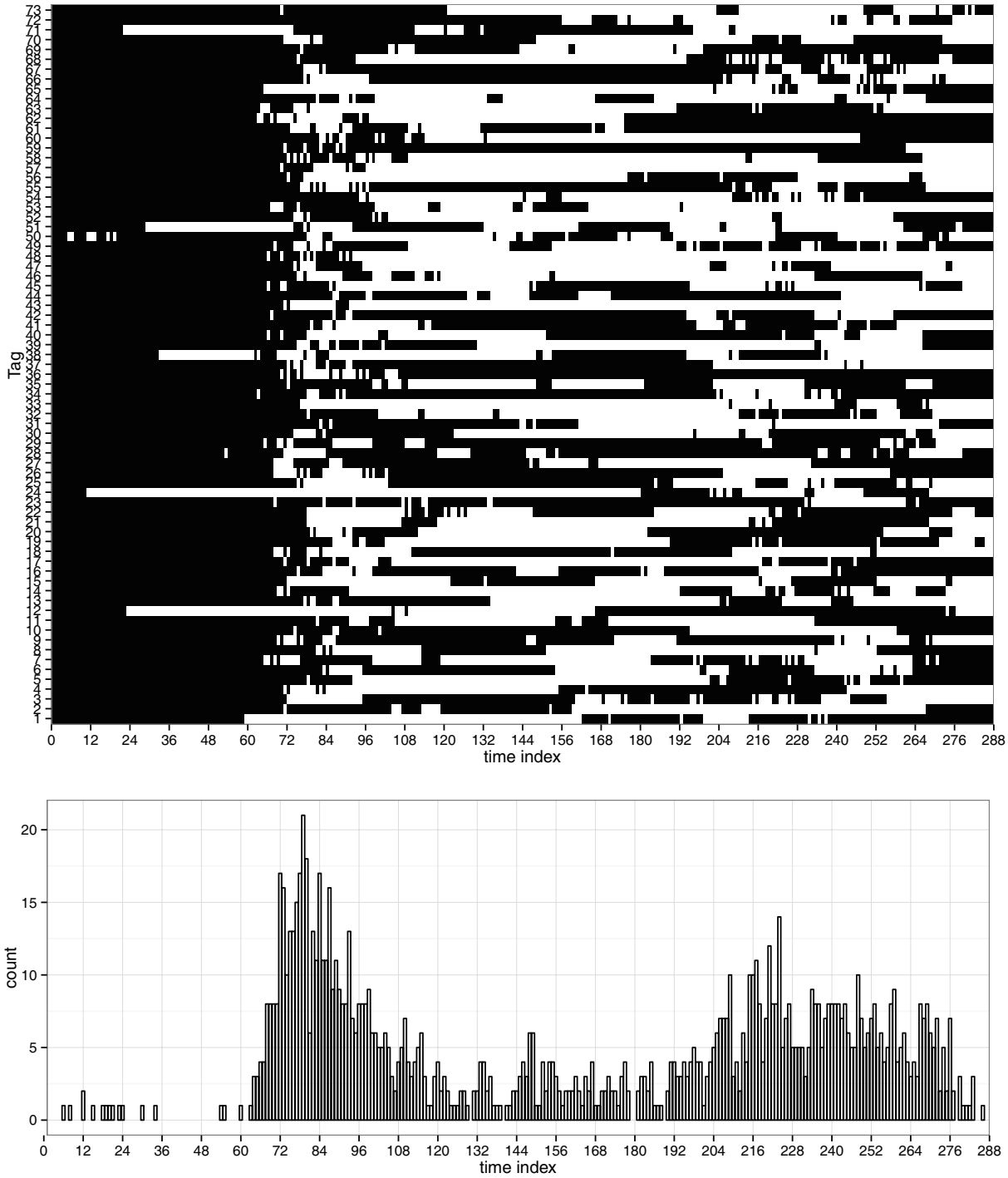


Figure 3.9 Weekday λ^{obs} defined slot cut. Top panel: x-axis is time sequence number from 1 to 288 (5 min per unit). y-axis is client tag from 1 to 73. For a client y, when the color changes at time index x, it means this client's consumption pattern is changing and we are moving to a new slot. Conversely, when the color remains the same, it means the process is considered to have a relatively stationary property, indicating that parameter estimation theory based on this stochastic assumption is applicable. Bottom panel: The bar-plot indicates how many clients possess a slot cut point at time index x.

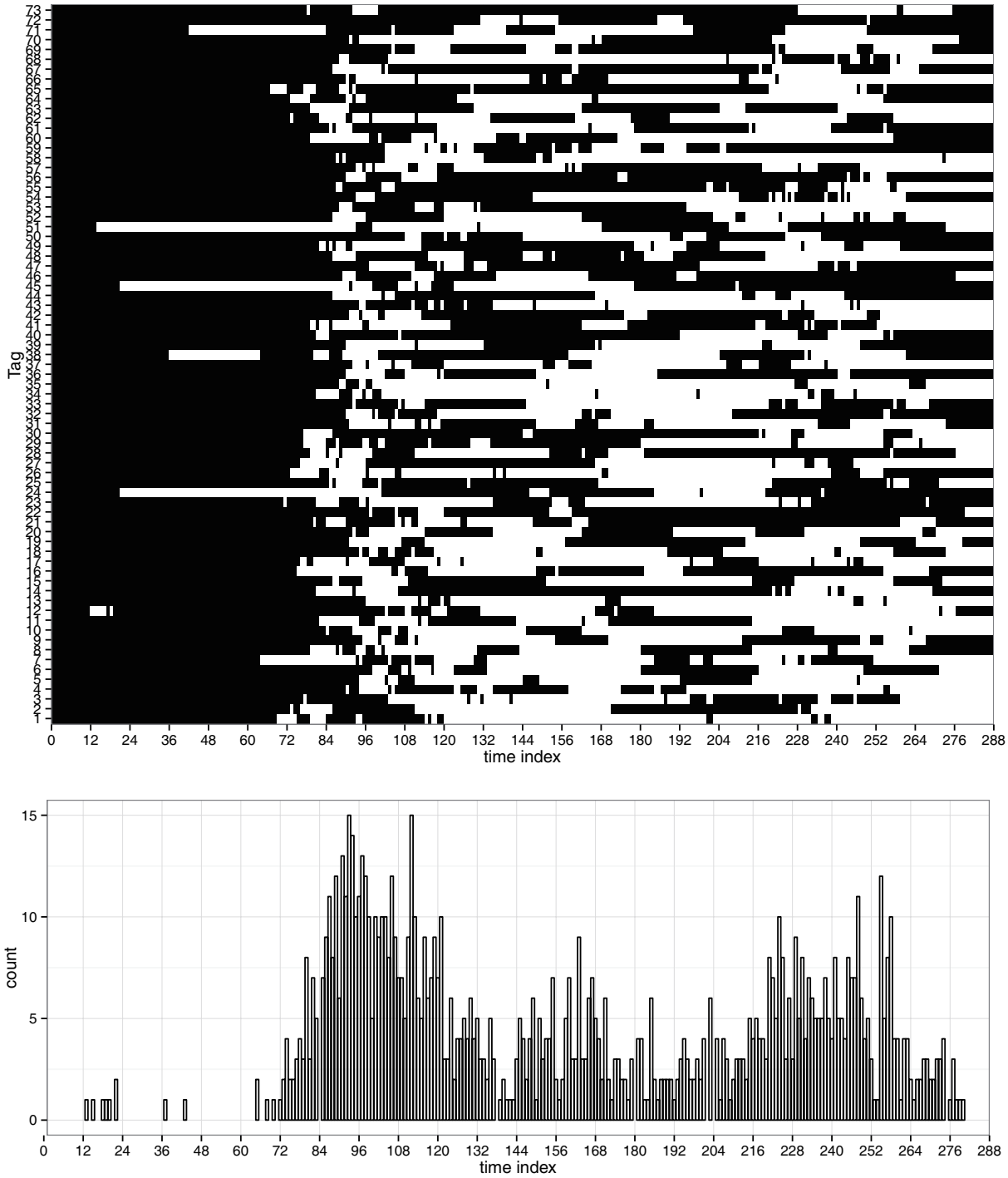


Figure 3.10 Weekend auto-selected λ model. Top panel: x-axis is time sequence number from 1 to 288 (5 min per unit). y-axis is client tag from 1 to 73. For a client y, when the color changes at time index x, it means this client's consumption pattern is changing and we are moving to a new slot. Conversely, when the color remains the same, it means the process is considered to have a relatively stationary property, indicating that parameter estimation theory based on this stochastic assumption is applicable. Bottom panel: The bar-plot indicates how many clients possess a slot cut point at time index x.

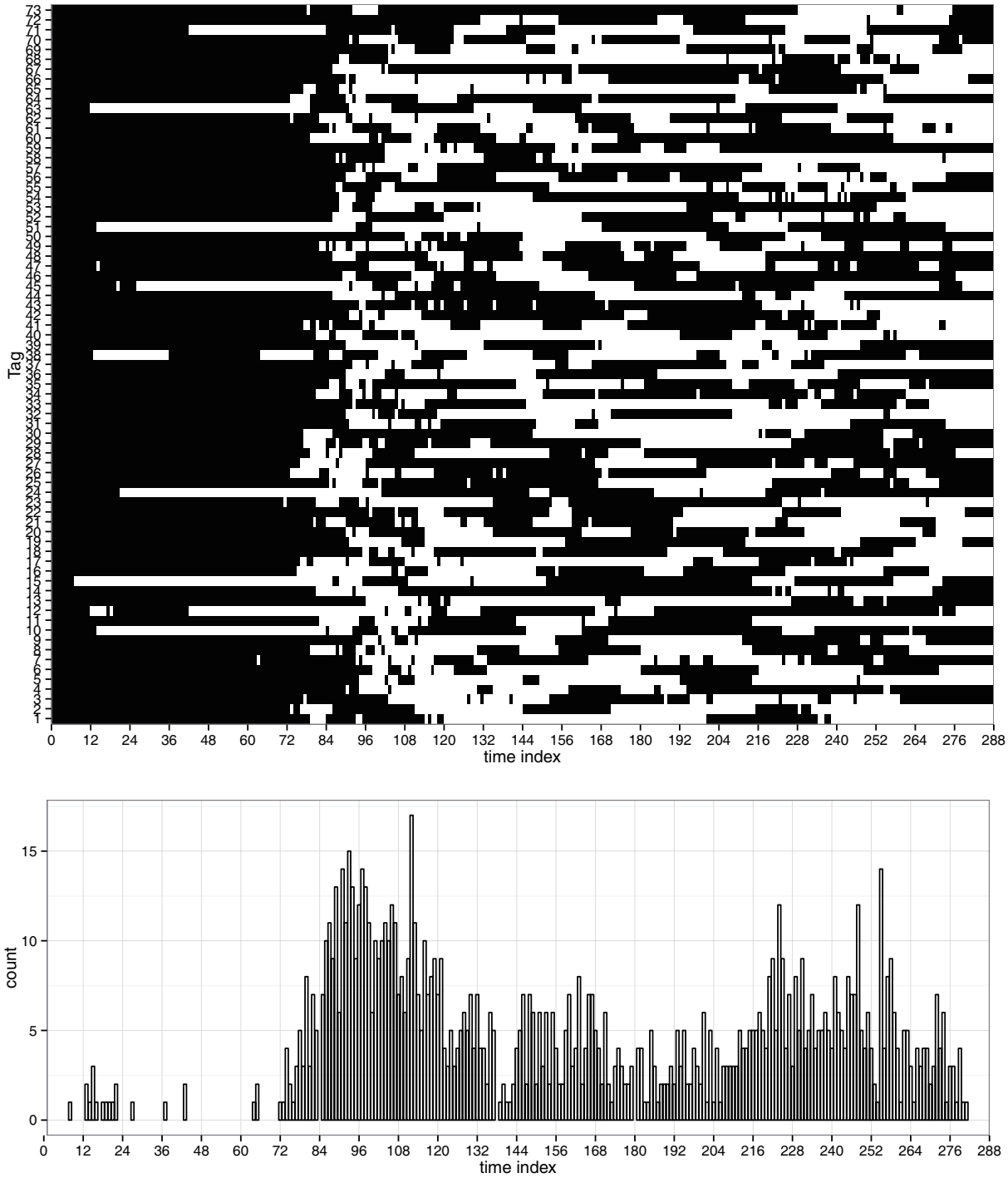


Figure 3.11 Weekend λ^{obs} defined slot cut. Top panel: x-axis is time sequence number from 1 to 288 (5 min per unit). y-axis is client tag from 1 to 73. For a client y, when the color changes at time index x, it means this client's consumption pattern is changing and we are moving to a new slot. Conversely, when the color remains the same, it means the process is considered to have a relatively stationary property, indicating that parameter estimation theory based on this stochastic assumption is applicable. Bottom panel: The bar-plot indicates how many clients possess a slot cut point at time index x.

CHAPTER 4 PARAMETER ESTIMATION

We have 73 customers' data in total. This section mainly presents the derivation of the method, which was used to estimate the water demand Markov chain time segment parameters for every single client on each defined time segment in Table 3.3. Indeed, the distribution of a hot water consumption event time durations was found to approximately follow an exponential distribution; and while off consumption periods may not be exponential, we find it analytically convenient to work with a two-state Markov chain water demand model.

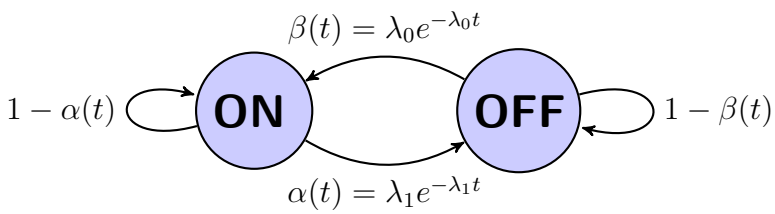


Figure 4.1 Two state Markov chain.

Figure 4.1 illustrates a two-state Markov model, where α and β are the transition probabilities from state "ON" to state "OFF" and from state "OFF" to state "ON" over time duration t. In this specified model, state "ON" (also called state 1) means consumption is present; its transition probability $\alpha(t)$ is the exponential distribution with rate λ_1, while state "OFF" means no consumption and is denoted state 0. Its transition probability is exponential with rate λ_0. Since in Section 3.2.3, a general slot segmentation has been decided, the non-stationary 24 hours process is divided into 6 time regions. It is assumed that consumption patterns are stationary within each of them. Therefore, the theory of stationary alternating renewal processes will be later applied to estimate from the data a triplet of constant parameters: "OFF" event rate λ_0 (/min), "ON" event rate λ_1 (/min) and average extraction rate c (L/min). Over each time homogeneous interval, the stochastic water demand process will be characterized by these three parameters, and three independent equations will need to be identified based on existing data to estimate these parameters. We choose to obtain these equations through a moment approach (El-Férik and Malhamé, 1994; Mortensen, 1990) with the following moments used: mean, variance, and autocovariance function at lag 1 (i.e. correlation between two successive measurements).

4.1 Alternating renewal process

Let $w(t)$ denote the continuous time alternating 1-0 renewal process of interest at time t with $f_1(\tau)$ and $f_0(\tau)$ the associated "ON" and "OFF" time durations stationary probability density functions (pdf's). Let $\mu_i = \int_0^\infty \tau f_i(\tau)d\tau$, $i = 0, 1$, be the expected "OFF" and "ON" durations. Assume that "ON" and "OFF" events duration time follow the exponential distributions. So

$$f_i(t) = \lambda_i e^{-\lambda_i t} \qquad i = 0, 1 \tag{4.1}$$

where $\lambda_i = 1/\mu_i$.

Also consider the water extraction rate c (L/min), when on, is a constant. Now the hot water consumption process is $z(t) = cw(t)$ while $Z(t) \triangleq \int_0^t z(\tau)d\tau = c\int_0^t w(\tau)d\tau = c\xi(t)$ is the total volume consumed random variable during an interval of length t, and $\xi(t)$ can be referred to as the *total busy time* during t. Given the fact that available customer data is hot water consumed over successive 5 min intervals, we shall develop expressions for the statistics of $Z_{t'}^t = \int_{t'}^{t'+t} z(\tau)d\tau$, $t = 5\text{min}$, $t' = Nt$, $N = 0, 1, 2, \ldots$.

4.2 Theoretical results

This section presents the moment method and develops the autocovariance functions under the equilibrium conditions in time domain and frequency domain.

4.2.1 Moments expressions

El-Férik and Malhamé have shown that the moments at stationarity of the Laplace transform of $\mathbb{E}[\xi(t)^n]$ process for a general alternating renewal process satisfy a relationship recursive in the power index n. In particular, based on their work, one obtains the following general formula for mean and variance[1]:

$$\mathbb{E}_s^{(\text{eq})}[Z(t)] = \frac{\mu_1 c}{\mu_0 + \mu_1}\frac{1}{s^2} \tag{4.2}$$

($\mathbb{E}^{(\text{eq})}[Z(t)] = \mu_1 ct/(\mu_0 + \mu_1)$ as expected) and

$$\mathbb{E}_s^{(\text{eq})}[Z(t)^2] = c^2\{\frac{\mu_1}{\mu_0 + \mu_1}\frac{2}{s^3} - \frac{2}{(\mu_0 + \mu_1)s^4}\frac{[1 - F_0(s)][1 - F_1(s)]}{[1 - F_0(s)F_1(s)]}\} \tag{4.3}$$

[1]The superscript $^{(\text{eq})}$ means the expectation is taken under equilibrium. The subscript $_s$ means the Laplace transform

where $F_i(s) = \mathcal{L}[f_i(t)] = 1/(1+\mu_i s) = \lambda_i/(\lambda_i + s)$, $i = 0, 1$.

4.2.2 Equilibrium autocovariance functions

Suppose $Z_0^t, Z_1^t, \ldots, Z_{(N-1)t}^t$ is a time series sequence of N observations; since the observation interval t is fixed at 5 min, we note simply $Z_{(i-1)t}^t = Z_i$ in the rest of this chapter. For a stationary stochastic process, the sample mean $\bar{Z}$ and variance $\hat{\sigma}_Z^2$ are given in Box et al. (2015)

$$\bar{Z} = \frac{1}{N}\sum_{i=1}^{N} Z_i \tag{4.4}$$

$$\hat{\sigma}_Z^2 = \frac{1}{N}\sum_{i=1}^{N}(Z_i - \bar{Z})^2 \tag{4.5}$$

Under the equilibrium conditions, the covariance between values Z_i and Z_{i+k}, separated by k intervals of time or by lag k, remains the same for all t. This autocovariance at lag k is

$$\gamma_k = \mathrm{cov}(Z_i, Z_{i+k}) = \mathbb{E}[(Z_i - \bar{Z})(Z_{i+k} - \bar{Z})] \tag{4.6}$$

Suppose X and Y are complete measure spaces. Suppose $f(x, y)$ is $X \times Y$ measurable. By Fubini's theorem, if

$$\int_{X\times Y} |f(x,y)| d(x,y) < \infty \tag{4.7}$$

where the integral is taken with respect to a product measure on the space over $X \times Y$, then

$$\int_X \left[\int_Y f(x,y)dy\right] dx = \int_Y \left[\int_X f(x,y)dx\right] dy = \int_{X\times Y} f(x,y)\, d(x,y) \tag{4.8}$$

the first two integrals being iterated integrals with respect to two measures, respectively, and the third being an integral with respect to a product of these two measures (Kudryavtsev, 2001).

Since the expectation of consumption integrated over a continuous time interval is always finite, it is written as

$$\mathbb{E}[Z(t)] = \mathbb{E}[\int_0^t z(\tau)d\tau] = \int_0^t \mathbb{E}[z(\tau)]d\tau \tag{4.9}$$

Then (4.6) becomes

$$\begin{aligned}\gamma_k &= c^2 \int_{it}^{(i+1)t} \int_{(i+k)t}^{(i+k+1)t} \mathbb{E}[w(\tau_1)w(\tau_2)]d\tau_1 d\tau_2 - c^2\bar{w} \int_{i}^{i+t} \int_{(i+k)t}^{(i+k+1)t} \mathbb{E}[w(\tau_1)] + \mathbb{E}[w(\tau_2)]d\tau_1 d\tau_2 \\ &+ c^2\bar{w}^2 \int_{it}^{(i+1)t} \int_{(i+k)t}^{(i+k+1)t} d\tau_1 d\tau_2 \\ &= c^2 \int_{it}^{(i+1)t} \int_{(i+k)t}^{(i+k+1)t} \mathbb{E}[w(\tau_1)w(\tau_2)]d\tau_1 d\tau_2 - c^2\bar{w}^2 t^2\end{aligned} \tag{4.10}$$

because $\mathbb{E}[w(\tau_1)] = \mathbb{E}[w(\tau_2)] = \bar{w}$ for a time invariant stochastic process at equilibrium and we also have $c^2\bar{w}^2t^2 = \bar{Z}^2$.

Mortensen first established a lemma in 1990 to compute the Laplace transform of $u(\tau_2 - \tau_1) = \mathbb{E}[w(\tau_1)w(\tau_2)]$ at equilibrium under stationary conditions (El-Férik and Malhamé, 1994; Mortensen, 1990)

$$U(s) = \mathcal{L}[u(\tau)] = \frac{\mu_1}{\mu_0 + \mu_1}\frac{1}{s} - \frac{1}{(\mu_0 + \mu_1)s^2}\frac{[1 - F_0(s)][1 - F_1(s)]}{[1 - F_0(s)F_1(s)]} \tag{4.11}$$

The Laplace transform of the equilibrium autocovariance function $\mathcal{L}[\gamma_k(t)]$ with window t are proven to be

$$\begin{cases} \dfrac{2c^2}{s^2}U(s) - \dfrac{2c^2\bar{w}^2}{s^3}, & k = 0, \\ -\dfrac{2c^2}{s^2}U(s) + \dfrac{2c^2}{s^2}U(\dfrac{s}{2}) - \dfrac{2c^2\bar{w}^2}{s^3}, & k = 1, \\ \dfrac{c^2(k-1)}{s^2}U(\dfrac{s}{k-1}) - \dfrac{2c^2k}{s^2}U(\dfrac{s}{k}) - \dfrac{c^2(k+1)}{s^2}U(\dfrac{s}{k+1}) - \dfrac{2c^2\bar{w}^2}{s^3} & k \geqslant 2. \end{cases} \tag{4.12}$$

in El-Férik and Malhamé (1994).

We use a multivariate function $H(s, m)$ to simplify the expression in (4.12)

$$H(s, m) = \frac{m}{s^2}U(\frac{s}{m}) = \frac{m^2}{s^2} \cdot \frac{1}{m}U(\frac{s}{m}) \tag{4.13}$$

with its Laplace inverse in continuous time

$$\begin{aligned} h(t, m) &= \mathcal{L}^{-1}[\frac{m^2}{s^2} \cdot \frac{1}{m}U(\frac{s}{m})] \\ &= \{\mathcal{L}^{-1}(\frac{m^2}{s^2}) * \mathcal{L}^{-1}[\frac{1}{m}U(\frac{s}{m})]\}(t) \\ &= m^2 \int_0^t u(m\tau)(t - \tau)d\tau \end{aligned} \tag{4.14}$$

(4.13), together with (4.12) yield:

$$\begin{cases} 2c^2H(s,m=1) - \dfrac{2c^2\bar{w}^2}{s^3}, & k=0, \\ -2c^2H(s,m=1) + c^2H(s,m=2) - \dfrac{2c^2\bar{w}^2}{s^3}, & k=1, \\ c^2H(s,m=k-1) - 2c^2H(s,m=k) - c^2H(s,m=k+1) - \dfrac{2c^2\bar{w}^2}{s^3} & k \geqslant 2. \end{cases} \tag{4.15}$$

alternatively in continuous time domain

$$\begin{cases} 2c^2h(t,m=1) - c^2\bar{w}^2t^2, & k=0, \\ -2c^2h(t,m=1) + c^2h(t,m=2) - c^2\bar{w}^2t^2, & k=1, \\ c^2h(t,m=k-1) - 2c^2h(t,m=k) - c^2h(t,m=k+1) - c^2\bar{w}^2t^2 & k \geqslant 2. \end{cases} \tag{4.16}$$

All of the above functions in the time domain or the frequency domain have expressions detailed in Table 4.1, in terms of the unknown statistical parameters λ_0, λ_1 and c.

4.3 Empirical estimate

It is indicated in Box et al. (2015) that the most satisfactory empirical estimate of the autocorrelation at lag k ρ_k is

$$r_k = \hat{\rho}_k = \frac{c_k}{c_0} \tag{4.17}$$

where c_k is the estimate of the autocovariance γ_k.

$$c_k = \hat{\gamma}_k = \frac{1}{N}\sum_{i=1}^{N-k}(Z_i - \bar{Z})(Z_{i+k} - \bar{Z}) \tag{4.18}$$

with $k = 0, 1, 2, \ldots, K$ and $K < N/4$ usually in practice. When $k = 0$, $\hat{\gamma}_0$ is the variance $\mathbb{V}[Z]$. The formulae for computing the empirical estimate value of $\mathbb{E}^{(\mathrm{eq})}[Z(t)]$, $\mathbb{E}^{(\mathrm{eq})}[Z(t)^2]$ and γ_k are listed in Table 4.1.

Given the data structure of sequences Z, which means N observations per day horizontally and multiple M days data observed vertically

$$Z = \begin{bmatrix} Z_{11} & Z_{12} & \cdots & Z_{1N} \\ Z_{21} & Z_{22} & \cdots & Z_{2N} \\ \vdots & \vdots & \ddots & \vdots \\ Z_{M1} & Z_{M2} & \cdots & Z_{MN} \end{bmatrix} \tag{4.19}$$

The calculation in (4.18) becomes

$$c_k = \hat{\gamma}_k = \frac{1}{MN} \sum_{j=1}^{M} \sum_{i=1}^{N-k} (Z_{ji} - \bar{Z})(Z_{j(i+k)} - \bar{Z}) \tag{4.20}$$

where $\bar{Z}$ is the overall average consumption value $\frac{1}{MN} \sum_{j=1}^{M} \sum_{i=1}^{N} Z_{ji}$.

On the other hand, we should be careful to use (4.17) and (4.20) only over the time homogeneous slots.

4.4 Estimation formulae

Given that there are three unknown λ_0, λ_1 and c, at least three equations are needed for them. So the simplest three cases are picked: the mean $\bar{Z}$, the variance $\gamma_0 = \mathbb{V}[Z]$ and the autocovariance at lag 1 γ_1. By referring to the terms $\mathbb{E}^{(\mathrm{eq})}[Z(t)]$ and γ_k $(k = 0, 1)$ in time domain and empirical estimate in Table 4.1, this system of three equations is developed as:

$$\begin{cases} \dfrac{\lambda_0 c t}{(\lambda_0 + \lambda_1)} = \hat{\bar{Z}} \\ \dfrac{2c^2\lambda_0\lambda_1 t}{(\lambda_0+\lambda_1)^3} + \dfrac{2c^2\lambda_0\lambda_1 e^{-(\lambda_0+\lambda_1)t}}{(\lambda_0+\lambda_1)^4} - \dfrac{2c^2\lambda_0\lambda_1}{(\lambda_0+\lambda_1)^4} = \hat{\gamma}_0 \\ -\dfrac{2c^2\lambda_0\lambda_1 e^{-(\lambda_0+\lambda_1)t}}{(\lambda_0+\lambda_1)^4} + \dfrac{c^2\lambda_0\lambda_1 e^{-(\lambda_0+\lambda_1)2t}}{(\lambda_0+\lambda_1)^4} + \dfrac{c^2\lambda_0\lambda_1}{(\lambda_0+\lambda_1)^4} = \hat{\gamma}_1 \end{cases} \tag{4.21}$$

Then the estimation work of the triplet $(\lambda_0, \lambda_1, c)$ for each client and slot is carried out in MATLAB.

4.5 Direct derivation of $\mathbb{E}_s^{(\mathrm{eq})}[w(r)w(r+t)]$ for exponential case

In what follows, we present an intuitive direct derivation of the expression of the Laplace transform of the equilibrium autocorrelation function $\mathbb{E}_s^{(\mathrm{eq})}[w(r)w(r+t)]$ whenever the ON-OFF durations of the alternating renewal process are exponential with respective distributions

$$f_i(\lambda_i, t) = \lambda_i e^{-\lambda_i t} \qquad i = 0, 1 \tag{4.22}$$

In the following subsection, we contrast our results with the general formula (Eq.3.29 in El-Férik and Malhamé (1994) and Eq.37 in Mortensen (1990)), when specialized to the exponential case, and show that although different at first sight, are in fact equivalent.

Table 4.1 Table of functions.

Term	**s domain**
$\mathbb{E}_s^{(\mathrm{eq})}[Z(t)]$	$\frac{\lambda_0 c}{\lambda_0+\lambda_1}\frac{1}{s^2}$
$\mathbb{E}_s^{(\mathrm{eq})}[Z(t)^2]$	$c^2[\frac{\lambda_0}{\lambda_0+\lambda_1}\frac{2}{s^3} - \frac{2\lambda_0\lambda_1}{(\lambda_0+\lambda_1)(\lambda_0+\lambda_1+s)s^3}]$
$U(s)$	$\frac{\lambda_0^2+\lambda_0 s}{(\lambda_0+\lambda_1)(\lambda_0+\lambda_1+s)s}$
$H(s,m)$	$\frac{m}{s^2}U(\frac{s}{m}) = \frac{m^2}{s^2}\cdot\frac{1}{m}U(\frac{s}{m})$
$\gamma_k(s)$	$\begin{cases} 2c^2H(s,m=1) - \dfrac{2c^2\bar{w}^2}{s^3}, & k=0, \\ -2c^2H(s,m=1) + c^2H(s,m=2) - \dfrac{2c^2\bar{w}^2}{s^3}, & k=1, \\ c^2H(s,m=k-1) - 2c^2H(s,m=k) - c^2H(s,m=k+1) - \dfrac{2c^2\bar{w}^2}{s^3} & k\geqslant 2. \end{cases}$
Term	**Time domain**
$\mathbb{E}^{(\mathrm{eq})}[Z(t)]$	$\frac{\lambda_0 c}{\lambda_0+\lambda_1}t$
$\mathbb{E}^{(\mathrm{eq})}[Z(t)^2]$	$c^2[\frac{\lambda_0^2}{(\lambda_0+\lambda_1)^2}t^2 + \frac{2\lambda_0\lambda_1}{(\lambda_0+\lambda_1)^3}t + \frac{2\lambda_0\lambda_1}{(\lambda_0+\lambda_1)^4}e^{-(\lambda_0+\lambda_1)t} - \frac{2\lambda_0\lambda_1}{(\lambda_0+\lambda_1)^4}]$
$u(t)$	$\frac{\lambda_0^2}{(\lambda_0+\lambda_1)^2} + \frac{\lambda_0\lambda_1}{(\lambda_0+\lambda_1)^2}e^{-(\lambda_0+\lambda_1)t}$
$h(t,m)$	$\frac{m^2\lambda_0}{(\lambda_0+\lambda_1)^2}[\frac{1}{2}\lambda_0 t^2 + \frac{\lambda_1}{(\lambda_0+\lambda_1)m}t + \frac{\lambda_1}{(\lambda_0+\lambda_1)^2m^2}e^{-(\lambda_0+\lambda_1)mt} - \frac{\lambda_1}{(\lambda_0+\lambda_1)^2m^2}]$
$\gamma_k(t)$	$\begin{cases} 2c^2h(t,m=1) - \bar{Z}^2, & k=0, \\ -2c^2h(t,m=1) + c^2h(t,m=2) - \bar{Z}^2, & k=1, \\ c^2h(t,m=k-1) - 2c^2h(t,m=k) - c^2h(t,m=k+1) - \bar{Z}^2, & k\geqslant 2. \end{cases}$
Term	**Empirical estimate**
$\mathbb{E}^{(\mathrm{eq})}[Z(t)]$	$\bar{Z}$
$\mathbb{E}^{(\mathrm{eq})}[Z(t)^2]$	$\mathbb{V}[Z] + \bar{Z}^2$
$\gamma_k(t)$	$\frac{1}{N}\sum_{i=1}^{N-k}(Z_i - \bar{Z})(Z_{i+k} - \bar{Z})$

The Laplace transform of these two exponential distributions pdf's are

$$F_0(s) = \mathcal{L}[f_0(t)] = \frac{\lambda_0}{\lambda_0 + s} \qquad F_1(s) = \mathcal{L}[f_1(t)] = \frac{\lambda_1}{\lambda_1 + s} \tag{4.23}$$

Also note the probability density function of k 1-0 ("ON" and "OFF") cycle durations as $f_{k,\mathrm{cyc}}(t)$ $(k = 1, 2, 3, \ldots)$ and its Laplace transform resulting from the Laplace transform of the convolution of k independent cycles is:

$$F_{k,\mathrm{cyc}}(s) = \mathcal{L}[f_{k,\mathrm{cyc}}(t)] = [F_0(s)F_1(s)]^k \tag{4.24}$$

4.5.1 Formula derivation of equilibrium autocovariance function

Let $\mathbb{E}[w(r)]$ denote the expected value of the $w(r)$ process at time r conditional on switching from 0 to 1 at time 0. We have by 0-1 renewal stationary process at equilibrium

$$\mathbb{E}^{(\mathrm{eq})}[w(r)] = \frac{\mu_1}{\mu_0 + \mu_1} = \frac{\lambda_0}{\lambda_0 + \lambda_1}. \tag{4.25}$$

In light of the memoryless property of exponential distribution, the equilibrium autocovariance function is

$$\mathbb{E}^{(\mathrm{eq})}[w(r)w(r+t)] = \mathbb{E}^{(\mathrm{eq})}[w(r)]\mathbb{E}[w(r+t) \mid w(r) = 1] \tag{4.26}$$

where $\mathbb{E}[w(r+t) \mid w(r) = 1]$ is the expected value with delay t given that the process is at state 1 at time r. During delay time t, there may appear one or more cycles or the process may always remain at state 1 on the interval $[r, r+t]$. So the expectation becomes a sum over all possible cases weighted by their respective probabilities.

$$\begin{aligned}\mathbb{E}^{(\mathrm{eq})}[w(r+t) \mid w(r) = 1] = \int_t^{+\infty} f_1(\tau)d\tau &+ \int_0^t f_{1,\mathrm{cyc}}(\tau) \int_{t-\tau}^{+\infty} f_1(\tau')d\tau' d\tau \\ &+ \int_0^t f_{2,\mathrm{cyc}}(\tau) \int_{t-\tau}^{+\infty} f_1(\tau')d\tau' d\tau \\ &+ \cdots\end{aligned} \tag{4.27}$$

Some transformations can be applied here

$$\int_t^{+\infty} f_1(\tau)d\tau = 1 - \int_0^t f_1(\tau)d\tau, \tag{4.28}$$

$$\int_{t-\tau}^{+\infty} f_1(\tau')d\tau' = 1 - \int_0^{t-\tau} f_1(\tau')d\tau'. \tag{4.29}$$

Thus, we have following expression for the equilibrium autocovariance function

$$\begin{aligned}\mathbb{E}^{(\mathrm{eq})}[w(r)w(r+t)] = \frac{\lambda_0}{\lambda_0 + \lambda_1}[1 &+ \int_0^t f_{1,\mathrm{cyc}}(\tau)d\tau + \int_0^t f_{2,\mathrm{cyc}}(\tau)d\tau + \cdots \\ &- (\int_0^t f_1(\tau)d\tau + \int_0^t f_{1,\mathrm{cyc}}(\tau) \int_0^{t-\tau} f_1(\tau')d\tau' d\tau \\ &+ \int_0^t f_{2,\mathrm{cyc}}(\tau) \int_0^{t-\tau} f_1(\tau')d\tau' d\tau + \cdots)].\end{aligned} \tag{4.30}$$

The next step is to take Laplace transform of (4.30). By the properties of Laplace transform, we have following terms

$$\mathcal{L}[\int_0^t f_{k,\text{cyc}}(\tau)d\tau] = \frac{F_{k,\text{cyc}}(s)}{s} = \frac{[F_0(s)F_1(s)]^k}{s}, \tag{4.31}$$

$$\mathcal{L}[\int_0^t f_1(\tau)d\tau] = \frac{F_1(s)}{s}, \tag{4.32}$$

$$\begin{aligned}\mathcal{L}[\int_0^t f_{k,\text{cyc}}(\tau) \underbrace{\int_0^{t-\tau} f_1(\tau')d\tau'}_{g(t-\tau)} d\tau] &= \mathcal{L}[f_{k,\text{cyc}}(\tau) * g(\tau)] \\ &= F_{k,\text{cyc}}(s) \cdot G(s) \\ &= [F_0(s)F_1(s)]^k \frac{F_1(s)}{s}.\end{aligned} \tag{4.33}$$

Note $\mathbb{E}_s^{(\text{eq})}[w(r)w(r+t)] = \mathcal{L}[\mathbb{E}^{(\text{eq})}[w(r)w(r+t)]]$, we have

$$\begin{aligned}\mathbb{E}_s^{(\text{eq})}[w(r)w(r+t)] &= \frac{\lambda_0}{\lambda_0+\lambda_1}\Big(\frac{1}{s} + \frac{F_0(s)F_1(s)}{s} + \frac{[F_0(s)F_1(s)]^2}{s} + \cdots \\ &\quad - \{\frac{F_1(s)}{s} + F_0(s)F_1(s)\frac{F_1(s)}{s} + [F_0(s)F_1(s)]^2\frac{F_1(s)}{s} + \cdots\}\Big) \\ &= \frac{\lambda_0}{\lambda_0+\lambda_1}\frac{1-F_1(s)}{s}\{1 + F_0(s)F_1(s) + [F_0(s)F_1(s)]^2 + \cdots\}.\end{aligned} \tag{4.34}$$

Since $|F_0(s)F_1(s)| = \left|\frac{\lambda_0}{\lambda_0+s}\frac{\lambda_1}{\lambda_1+s}\right| < 1$ for Re(s)>0, the above sums converge, and (4.34) becomes

$$\mathbb{E}_s^{(\text{eq})}[w(r)w(r+t)] = \frac{\lambda_0}{\lambda_0+\lambda_1}\frac{1-F_1(s)}{s[1-F_0(s)F_1(s)]}. \tag{4.35}$$

4.5.2 Comparison with general alternating renewal process formula

As reported in (4.11), the Laplace transform of the equilibrium autocovariance function of the $w(r)$ process is given by:

$$U(s) = \mathcal{L}[u(t)] = \frac{\mu_1}{\mu_0+\mu_1}\frac{1}{s} - \frac{1}{(\mu_0+\mu_1)s^2}\frac{[1-F_0(s)][1-F_1(s)]}{[1-F_0(s)F_1(s)]}, \tag{4.36}$$

See Eq.3.29 in El-Férik and Malhamé (1994) and Eq.37 in Mortensen (1990).

The expansion form of (4.36) with $f_0(t)$, $f_1(t)$ probability density functions of exponential

distributions is

$$
\begin{aligned}
U(s) &= \frac{\lambda_0}{\lambda_0+\lambda_1}\frac{1}{s} - \frac{\lambda_0\lambda_1}{(\lambda_0+\lambda_1)s^2}\cdot\frac{[1-F_0(s)][1-F_1(s)]}{[1-F_0(s)F_1(s)]} \\
&= \frac{\lambda_0}{\lambda_0+\lambda_1}\frac{1}{s[1-F_0(s)F_1(s)]}\{1-F_0(s)F_1(s)-\frac{\lambda_1}{s}[1-F_0(s)-F_1(s)+F_0(s)F_1(s)]\} \\
&= \frac{\lambda_0}{\lambda_0+\lambda_1}\frac{1}{s[1-F_0(s)F_1(s)]} \\
&\quad [\frac{s(\lambda_0+s)(\lambda_1+s)-\lambda_0\lambda_1 s-\lambda_1(\lambda_0+s)(\lambda_1+s)+\lambda_0\lambda_1(\lambda_1+s)+\lambda_1^2(\lambda_0+s)-\lambda_0\lambda_1^2}{s(\lambda_0+s)(\lambda_1+s)}] \\
&= \frac{\lambda_0}{\lambda_0+\lambda_1}\frac{1}{s[1-F_0(s)F_1(s)]}\frac{1}{\lambda_1+s} \\
&= \frac{\lambda_0}{\lambda_0+\lambda_1}\frac{1-F_1(s)}{s[1-F_0(s)F_1(s)]}. \qquad (4.37)
\end{aligned}
$$

The result in (4.37) concurs with that obtained in (4.35). Thus, formula (4.35) obtained by direct arguments valid only for the special case of two state Markov chains, is consistent with the general formula obtained by both Mortensen (1990) and El-Férik and Malhamé (1994).

4.6 Simulation

The proposed parameter estimation method in this chapter is verified through simulation in subsection 4.6.1, while subsection 4.6.2 provides a specific client example. The simulation discrepancies relative to observed data are discussed in subsection 4.6.3.

4.6.1 Simulation validation of method

An assumption has been made that processes are considered to be stationary on each slot, but this is not always the truth in reality. Even if the best general slot segmentation is chosen for all clients, the individual may still have a non-stationary consumption pattern. However, predicting the whole quantity of consumption is more of our interest rather than predicting the exact peak/valley occurring time in one slot. So when the method is applied to a quasi-stationary process and the parameters $(\lambda_0, \lambda_1, c)$ have been estimated for a client on a slot, we use them to run a simulation of stationary 0-1 alternating renewal process and compare the $(\bar{Z}, \gamma_0, \gamma_1)_{\text{sim}}$ of simulated data with $(\bar{Z}, \gamma_0, \gamma_1)_{\text{obs}}$ of observed raw data. If they match with each other, the proposed method is verified.

The simulated data has the same sequence length N and number of observed days M as the raw data. The "ON" and "OFF" events' duration time are randomized from the distributions $\exp(\lambda_1)$ and $\exp(\lambda_0)$. Another parameter to set is the system initial state, "ON" or "OFF", at

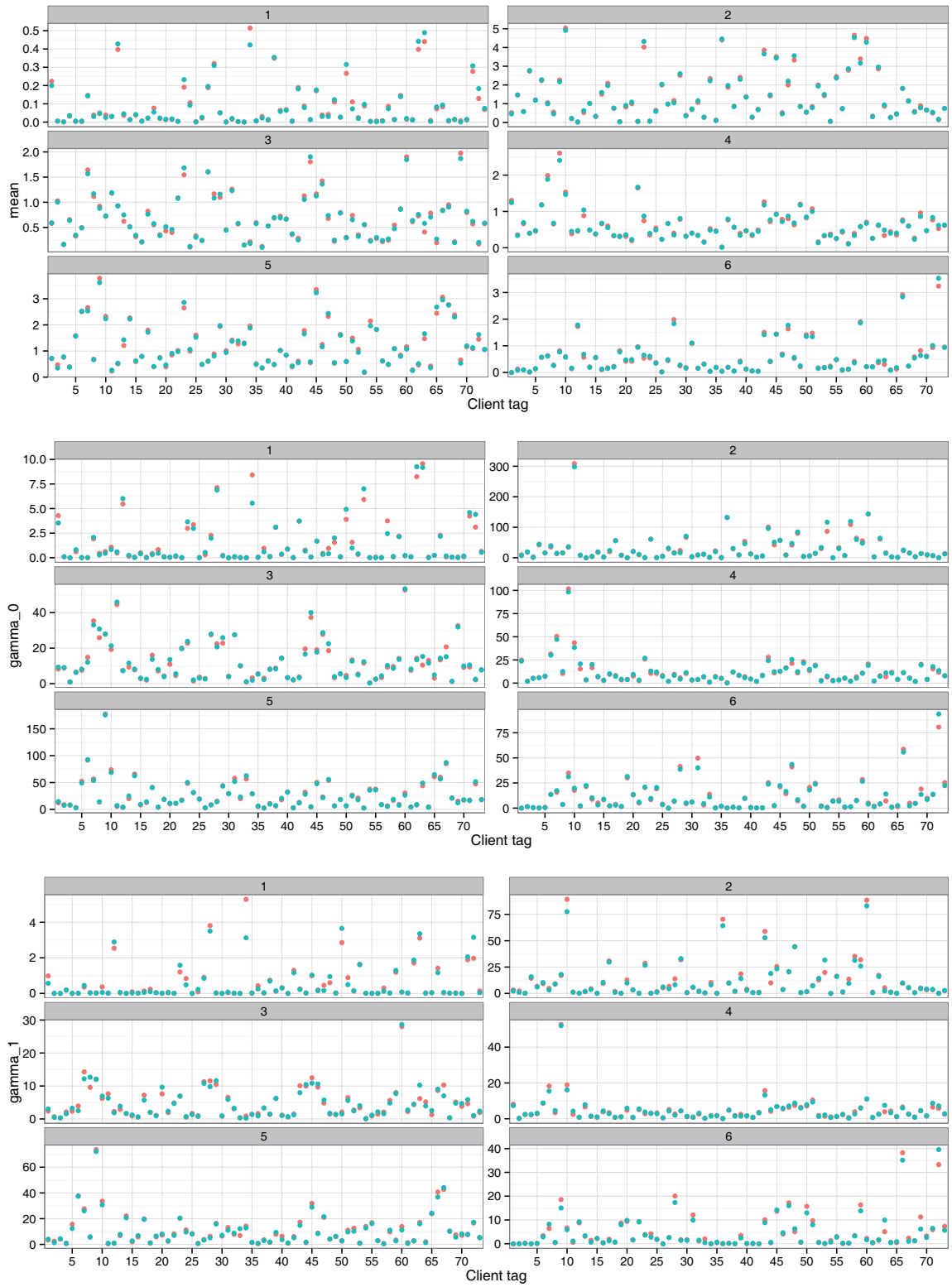


Figure 4.2 Observed (red) and simulation (blue) values of $\hat{\bar{Z}}$, $\hat{\gamma}_0$, $\hat{\gamma}_1$ on weekday. x-axis indicates the client tag, while y-axis are the values. The slot numbers are marked on the top of each facet.

the beginning of each day m simulated. So a Bernoulli test is launched with the probability of having state "ON" as the initial state $p_{ON}^{\text{initial}} = (1/\lambda_0)/(1/\lambda_0 + 1/\lambda_1) = \lambda_1/(\lambda_0 + \lambda_1)$. After simulating the alternative events with exponential distributions defined above, the 5 min windows are taken to segment time series data and calculate occupation time ξ_i during the i-th 5 min unit time. Then volume consumed $Z_i = c\xi_i$ because the extraction rate is assumed to be a constant. Plus, the first 48 5 min unit time data Z_i are abandoned. This is to offer a long enough period to warm up before observing. Then only the data collected after this offset time is taken for computing empirical values $(\bar{Z}, \gamma_0, \gamma_1)_{\text{sim}}$.

Figure 4.2 illustrates in weekday scenario, the comparison of the mean $\bar{Z}$ (top), the variance γ_0 (middle) and the autocovariance at lag 1 γ_1 (bottom) between the observed (red) and simulation (blue) ones. This figure shows that simulation results of different clients are given to illustrate the proposed method, since most observed values match well with the simulated ones.

4.6.2 Single client simulation

In this section, a visualization of comparison between simulated and observed data is given with special regard to client 6 on weekdays (Figure 4.3). The observed raw data is shown on the top of this figure and the bottom one represents the simulation.

The simulation part looks very similar to the observed one, except that the consumption events are more spread out over the entire slot region, because we are modeling it with a stationary process. When there is an obvious peak time, for example, regarding the second slot between time index 84 and 96, the simulation cannot reflect this information to us. However, as it has been mentioned before, whether the whole quantity of client's consumption of one slot can be well predicted is more concerned. Figure 4.4 shows client 6's average consumption over time on the 114 days, the sum of the black and the sum of the red points on each slot, which means the whole consumption of observed and simulated data, tend to be very close.

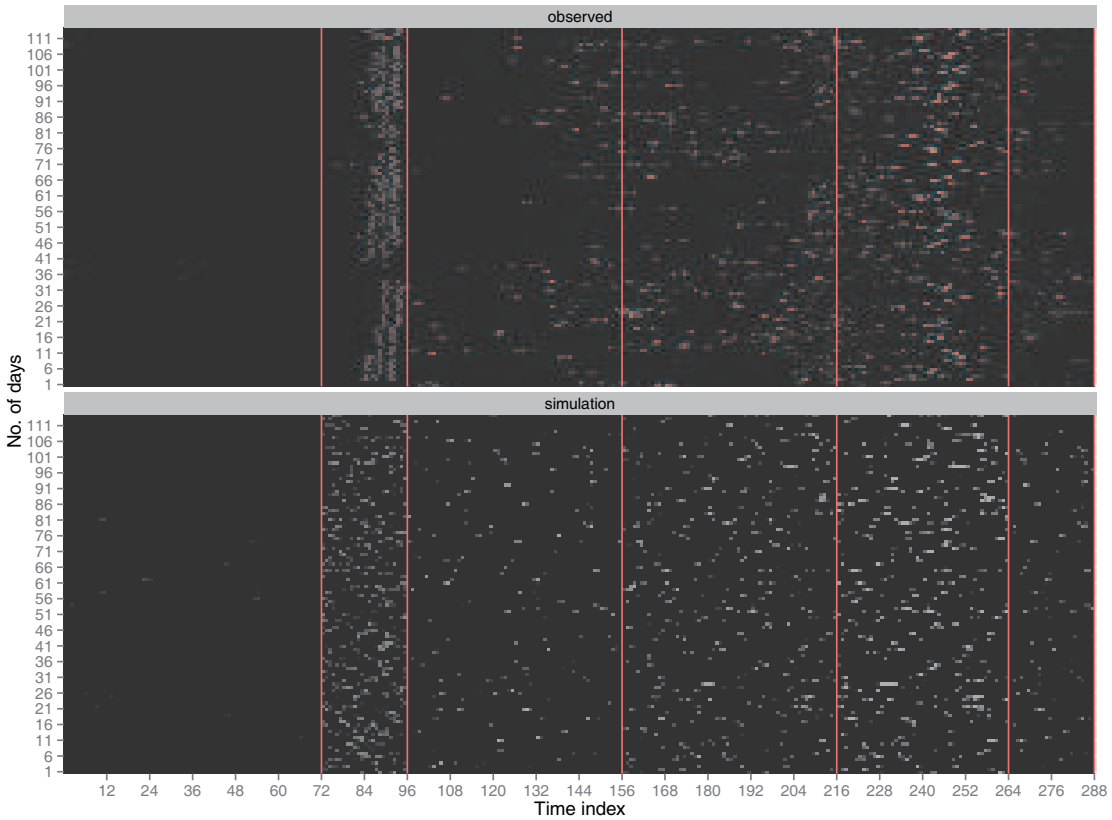


Figure 4.3 Observed (top) and simulation (bottom) data of client 6 on weekday. Client 6 has 114 days observed data on weekdays and they are arranged vertically on y-axis. x-axis shows the 288 time index of 24 hours during one day and the six different slots are separated by red lines. The status of consumption at time index x of day y is shown in colors. Black means there is no consumption, while where the gray scale is higher means a larger quantity of consumption is present.

4.6.3 Error by slots for all clients

After the discussion on a specific client, Figure 4.5 shows the histograms of consumption slot error between observed data and simulation using proposed method on weekday and on weekend scenarios. Most of the cases' errors are centred around 0 and minority of them are beyond 10L/slot. Considered that the minimum slot length is 2 hours, the range of errors is acceptable.

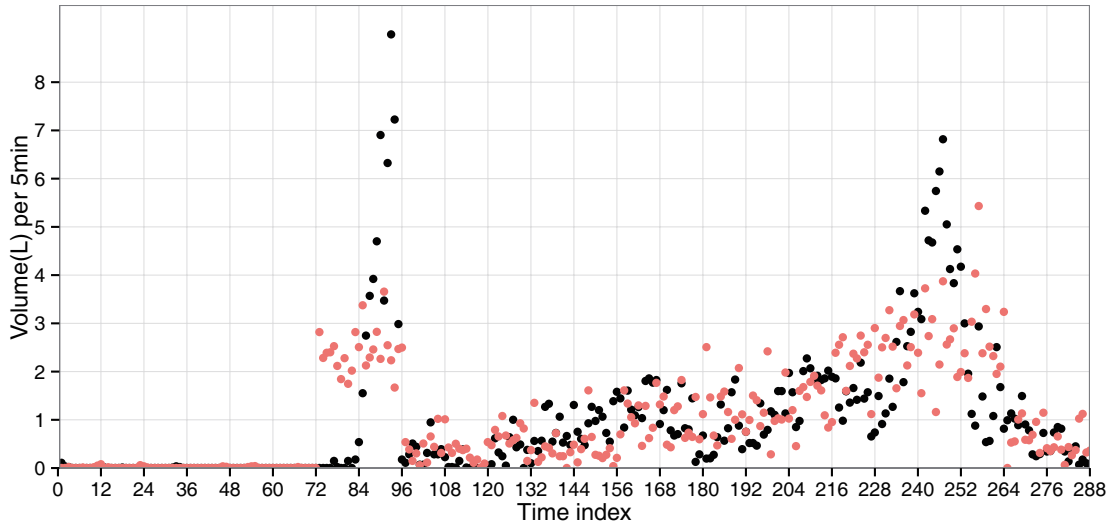


Figure 4.4 Average observed (black) and simulation (red) data of client 6 on weekday. x-axis shows the 288 time index of 24 hours during one day and y-axis is the average consumption over the client 6's 114 weekdays data.

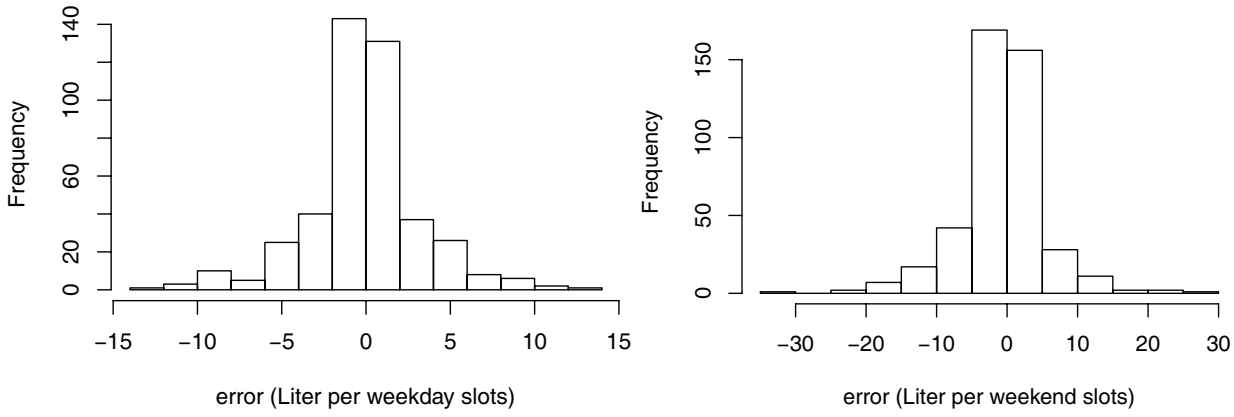


Figure 4.5 Histogram of slot simulation discrepancies relative to observed data on weekday (left) and weekend (right). In each of weekday/weekend scenarios, 73 clients $\times$ 6 slots $=$ 438 cases are counted.

CHAPTER 5 CLUSTERING RESULTS

While chapter 4 explains how λ_0, λ_1 and c are estimated from the data for each client on each slot, this chapter presents a "bottom-up" work, which means on each slot, all clients' behaviour are grouped into different aggregated sub-populations using this triplet. These three features are chosen because λ_0 and λ_1 indicate the arrival rates of "ON" and "OFF" events and c is the extraction rate, reflecting the level of demand of the client. For a two-state stationary process, the consumption pattern is adequately characterized by these parameters.

Figure 5.1 helps to understand the two steps at this stage. The first clustering is on only λ_0 and λ_1, since we want to find the clients who share the similar time frequency cycle pattern in the same slot. Then within each of the clusters, a sub-clustering on the extraction rate c is taken to distinguish demand levels into groups. The dashed line nodes indicate there are multiple cases. The directions of the arrows point to subdivided cases.

Finite Gaussian mixture modeling, is applied by the R package **mclust** version 5.2 using the EM algorithm, which is a powerful and wildly used clustering method (Scrucca et al., 2016). This clustering method is explained in more details in Section 5.1. The results and the implications are presented in Section 5.2.

5.1 Clustering method

The Gaussian mixture modeling technique applied in this thesis is presented in detail in this section. One of the advantages of Gaussian mixture model (GMM) is that it not only assigns each data point to one of the clusters but also gives the probability that these data points are assigned to each cluster. It learns the Gaussian distribution parameters for each cluster, which would be useful in simulating large population's behaviour. Moreover, with GMM, clusters can have different sizes and correlation structures within them.

It can be postulated that the customers are drawn independently from several distributions with different hot water consumption pattern. We are going to cluster their estimated rates $(\hat{\lambda}_0, \hat{\lambda}_1)$ and $\hat{c}$. From the point of view of the central limit theorem, if the number of customers approaches infinity, the estimated rates will converge in distribution to (multivariate) Gaussian distributions. So it is reasonable to assume an heterogeneous model as a mixture of Gaussian distributions. Therefore, GMM clustering can be more appropriate to use than, e.g, k-means clustering.

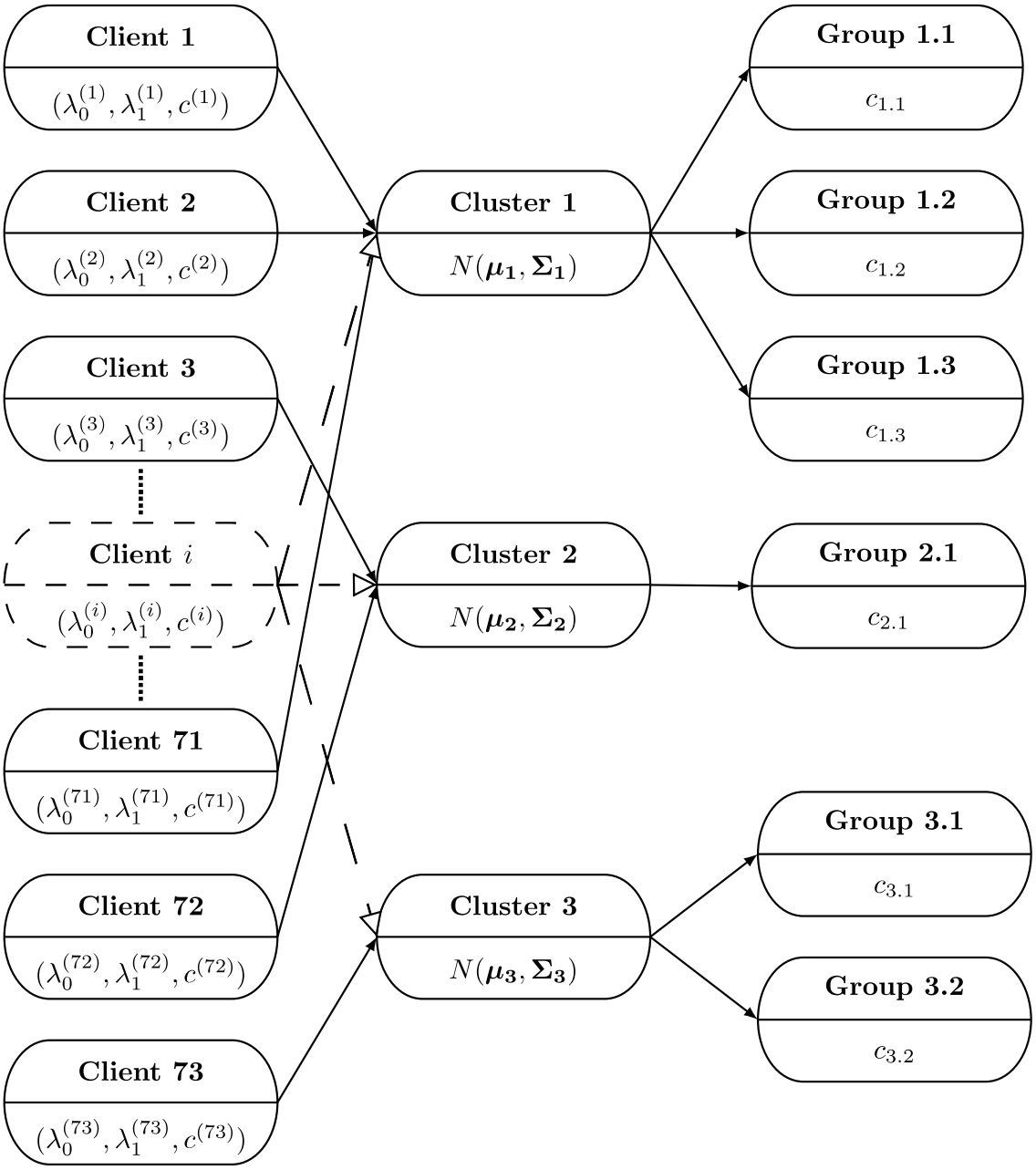


Figure 5.1 Methodology diagram part 2. The first clustering is on only λ_0 and λ_1, since we want to find the clients who share the similar time frequency cycle pattern in the same slot. Then within each of the clusters, a sub-clustering on the extraction rate c is taken to distinguish demand levels into groups. The superscript $^{(i)}$ indicates the client tag i. The dashed line nodes indicate there are multiple cases. The directions of the arrows point to subdivided cases.

5.1.1 Mixture model

Mixture model assumes that the overall population consists of a finite number of sub-populations. The data density is

$$f(\boldsymbol{x}_i; \boldsymbol{\Psi}) = \sum_{k=1}^{G} \pi_k f_k(\boldsymbol{x}_i; \boldsymbol{\theta}_k), \tag{5.1}$$

where a sample of n observations $\boldsymbol{x} = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n\}$ is independent and identically distributed by G components, with $f_k(\boldsymbol{x}_i; \boldsymbol{\theta}_k)$, $k = 1, \ldots, G$, the density function of component k for observation $\boldsymbol{x}_i$ with mixing probability π_k ($\sum_{k=1}^{G} \pi_k = 1$) and parameter vector $\boldsymbol{\theta}_k$. $f(\boldsymbol{x}_i; \boldsymbol{\Psi})$ is the mixture density function with unknown statistical parameter vector $\boldsymbol{\Psi} = \{\pi_1, \ldots, \pi_{G-1}, \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_G\}$.

5.1.2 Expectation-maximization (EM) algorithm

The mixture model is often used in model-based unsupervised learning for estimating the unknown $\boldsymbol{\Psi}$ and partitioning the observations $\boldsymbol{x}$ into meaningful sub-populations. The log-likelihood function of (5.1) is $\ell(\boldsymbol{\Psi}; \boldsymbol{x}) = \sum_{i=1}^{n} \log \{f(\boldsymbol{x}_i; \boldsymbol{\Psi})\}$. It is hard to acquire the direct maximizer of $\ell(\boldsymbol{\Psi}; \boldsymbol{x})$ because of the sum of logarithmic terms. The expectation-maximization (EM) algorithm is a popular tool to find (locally) maximum likelihood parameters. A vector of binary indicator variables $\boldsymbol{z} = (\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n)$, where $\boldsymbol{z}_i = (z_{i1}, \ldots, z_{iG})$, is called missing data. z_{ik} is the membership indicator

$$z_{ik} = \begin{cases} 1 & \text{if } \boldsymbol{x}_i \text{ belongs to component } k \\ 0 & \text{otherwise} \end{cases} \tag{5.2}$$

with condition $\sum_{k=1}^{G} z_{ik} = 1$. Then the log-likelihood of complete data $(\boldsymbol{x}, \boldsymbol{z})$ is

$$\ell_0(\boldsymbol{\Psi}; \boldsymbol{x}, \boldsymbol{z}) = \sum_{i=1}^{n} \sum_{k=1}^{G} \gamma_{ik}(\boldsymbol{\Psi}) \log \left(\pi_k p_{\boldsymbol{x}|\boldsymbol{z}}(\boldsymbol{x}_i | \boldsymbol{\Psi}_k)\right) \tag{5.3}$$

where the "responsibilities" are given by $\gamma_{ik}(\boldsymbol{\Psi}) = \mathbb{E}[z_{ik}|\boldsymbol{\Psi}, \boldsymbol{x}_i] = p(z_{ik} = 1|\boldsymbol{\Psi}, \boldsymbol{x}_i)$. The responsibility $\gamma_{ik}(\boldsymbol{\Psi})$ stands for the probability that $\boldsymbol{x}_i$ is generated by component k given $\boldsymbol{\Psi}$ and $\boldsymbol{x}_i$.

EM algorithm takes iterative steps (Dempster et al., 1977; Wu, 1983; Friedman et al., 2001):

1. Take initial guesses for parameters $\hat{\boldsymbol{\Psi}}^{(0)}$.

2. *Expectation step*: at the j-th step, compute $Q(\boldsymbol{\Psi}, \boldsymbol{\Psi}') = \mathbb{E}_{\boldsymbol{z}|\boldsymbol{x}, \boldsymbol{\Psi}'}[\ell_0(\boldsymbol{\Psi}; \boldsymbol{x}, \boldsymbol{z})]$ as a func-

tion of dummy argument $\boldsymbol{\Psi}'$.

3. *Maximization step*: find the new estimate $\hat{\boldsymbol{\Psi}}^{(j+1)} = \underset{\boldsymbol{\Psi}}{\operatorname{argmax}}\, Q(\boldsymbol{\Psi}, \hat{\boldsymbol{\Psi}}^{(j)})$ through the maximum likelihood estimator (MLE) of $Q(\boldsymbol{\Psi}, \hat{\boldsymbol{\Psi}}^{(j)})$ over $\boldsymbol{\Psi}$.

4. Iterate steps 2 and 3 until convergence.

5.1.3 mclust and mixture modeling

If in (5.1) all the G components are multivariate Gaussian distributions $N(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \quad (\forall k = 1, 2, \ldots, G)$, the mixture is called Gaussian mixture model (GMM).

Note that the clients of different behaviour patterns form a heterogeneous population, but they can also be grouped into typical sub-populations. In this research, the clients are assumed to be independent and identically distributed by several Gaussian distributions, from which the central limit theorem follows as the number of clients grows to infinity. For the 73 clients in each time segment on weekdays or weekends, first a bivariate Gaussian mixture for the birth and termination rates of water events $(\lambda_0^{(i)}, \lambda_1^{(i)}) \sim N(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$ is modeled to find the clusters of different time frequency cycle pattern. Then, within each cluster, the members are subdivided into several groups using the univariate Gaussian mixture modeling $c^{(i)} \sim N(\mu_k, \sigma_k)$ to further distinguish the demand levels according to the rate of extraction c. Figure 5.1[1] helps to understand these steps. The finite Gaussian mixture modeling is realized by the R package **mclust** using the EM algorithm.

For multivariate Gaussian mixture model, volume, shape and orientation of within-group covariance $\boldsymbol{\Sigma}_k$ of density contours centred at $\boldsymbol{\mu}_k$ are constrained to be equal or to be variable across components. There are 14 possible models with different geometric characteristics for multidimensional data in the **mclust** package. For univariate Gaussian mixture model, it provides 2 models depending on assumptions over the variance and the mean for each cluster (Banfield and Raftery, 1993; Celeux and Govaert, 1995; Scrucca et al., 2016).

The EM algorithm is used to find the maximum likelihood estimator. Furthermore, **mclust** uses the Bayesian information criterion[2] (BIC; Schwarz, 1978) to perform model selection. It is an asymptotic result derived under the assumptions that the data distribution is in the exponential family. It penalizes the complexity of the model by introducing a penalty term for the number of parameters. The BIC addresses two questions in GMMs: (i) select the model structure $\mathcal{M}$ among the ones described above (ii) determine the appropriate number

[1]The superscript $^{(i)}$ in Figure 5.1 accounts for the client i or in other words the i-th observation.

[2]In the **mclust** package, the BIC is defined as (5.4), which is opposite to its definition by default (Fraley and Raftery, 1998).

of Gaussian distribution components G.

$$\mathrm{BIC}_{\mathcal{M},G} \equiv 2\ell_{\mathcal{M},G}(\boldsymbol{x}|\hat{\boldsymbol{\Psi}})) - v\log(n) \tag{5.4}$$

where $\ell_{\mathcal{M},G}(\boldsymbol{x}|\hat{\boldsymbol{\Psi}})$ is the log-likelihood at the MLE $\hat{\boldsymbol{\Psi}}$ for model $\mathcal{M}$ with G components, n is the sample size, and v is the number of estimated parameters. The larger the value of the $\mathrm{BIC}_{\mathcal{M},G}$ is, the stronger evidence the model of type $\mathcal{M}$ with G components has. (Fraley and Raftery, 1998; Scrucca et al., 2016).

There are some other measures of model selection, such as the Akaike information criterion (AIC; Akaike, 1970). The AIC is defined as (5.4) with the penalty term $\log(n)$ replaced by the factor 2. Although they share the similar form, they are derived from different perspectives. The AIC is oriented from the concept of entropy in information theory, while the BIC is motivated by maximizing the posterior model probability. In general, the BIC is consistent whereas the AIC is not so. In other words, when $n \to \infty$, the probability of selecting the true model in the model space by BIC approaches 1. The BIC's high penalty on complexity helps for identification when the true model is simple or finite-dimensional. If the true model is complex and infinite-dimensional, the AIC would be a better choice (Shao, 1997). Given that we are modeling a finite Gaussian mixture model, the BIC is shown to be optimal.

It follows that the final grouping described by the Gaussian distribution components are the ultimate objective of this study, i.e. describe the sub-population hot water consumption pattern.

5.2 Clustering results

Table 5.3, Table 5.4 and Table 5.5 are the final clustering results of the 73 clients on weekday and weekend scenarios. They are explained by the following two subsections with examples.

5.2.1 Clustering on λ_0 and λ_1

The clustering on λ_0 and λ_1 is applied in each time segment. In bivariate case, for the k-th estimated Gaussian distribution component $N(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$ or in other word the cluster k, the mean $\boldsymbol{\mu}_k$ and covariance matrix $\boldsymbol{\Sigma}_k$ are

$$\boldsymbol{\mu}_k = \begin{pmatrix} \mu_{\lambda_0} \\ \mu_{\lambda_1} \end{pmatrix}_k \quad \boldsymbol{\Sigma}_k = \begin{pmatrix} \sigma^2_{\lambda_0} & \rho\sigma_{\lambda_0}\sigma_{\lambda_1} \\ \rho\sigma_{\lambda_0}\sigma_{\lambda_1} & \sigma^2_{\lambda_1} \end{pmatrix}_k \tag{5.5}$$

where μ_{λ_i} $(i = 0, 1)$ the mean of λ_i, σ_{λ_i} is the standard deviation of the Gaussian distribution component estimated in GMM, ρ the correlation between λ_0 and λ_1. The standard error of the mean (SEM), which quantifies uncertainty to estimate the mean, is related to the standard deviation of the distribution by:

$$SEM_{\lambda_i} = \frac{\sigma_{\lambda_i}}{\sqrt{n}} \tag{5.6}$$

where n is the number of clients within the cluster.

The clustering results are shown on the left side in Table 5.3, 5.4 and 5.5 by

Slot.no	indicator of time segments defined in Table 3.3, from 1 to 6
Cluster	indicator of clusters[3]
Number of clients	number of clients contained in each cluster
μ_{λ_0}	the mean of the birth rate of hot water events λ_0 for each cluster
SEM_{λ_0}	the standard error of the mean of λ_0 for each cluster
μ_{λ_1}	the mean of the termination rate of hot water events λ_1 for each cluster
SEM_{λ_1}	the standard error of the mean of λ_1 for each cluster
ρ	the correlation between λ_0 and λ_1 for each cluster

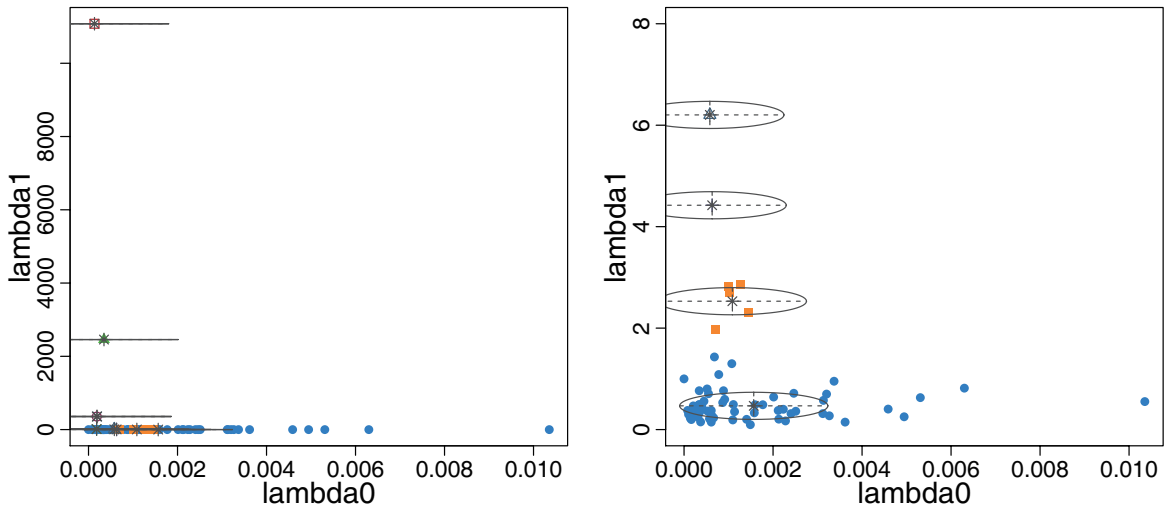
The asymptotic 95% confidence interval is defined as

$$\hat{\mu} \pm 1.96 \times SEM \tag{5.7}$$

According to the values shown in Table 5.3, 5.4 and 5.5, the mean μ_{λ_0}, μ_{λ_1} are generally well estimated. But when cluster size is relatively small, the uncertainty increases. For instance, cluster 2 of 8 clients on slot 2 of weekday scenario and cluster 2 of 8 clients on slot 1 of weekend's.

An example of clustering on (λ_0, λ_1) of slot 1 on weekday scenario is presented in Figure 5.2. According to Table 3.3, slot 1 on weekday represents the time interval from 0h00 to 6h00 in the morning. Usually it is a silent period since people are sleeping and less water consumption presents during this time. So it is reasonable to have small λ_0 and large λ_1. Furthermore, several clients have extremely large value for λ_1 because of no consumption observed in early morning. They are considered as outliers in final results on account of its small cluster size.

[3]When a cluster size is less than 8, the elements contained are counted as "outliers" and are excluded from the final results. Because the population size is not large enough to regard it as a typical type.



(a) (b)

Figure 5.2 Clustering result of slot 1 on weekday. (a) overall view λ_0 versus λ_1 of slot 1 on weekday. (b) zoom on y-axis from 0 to 8 of (a). **mclust** plots the data points in different colors to distinguish the clusters found. For each Gaussian distribution component fitted, an ellipse centred at $(\mu_{\lambda_0}, \mu_{\lambda_1})$ with semi-major axis σ_{λ_0} and semi-minor axis σ_{λ_1} is plotted.

5.2.2 Clustering c

Within each cluster found according to λ_0 and λ_1, a sub-cluster on the extraction rate c (L/min) is taken. The clustering results are shown on the right side in Table 5.3, 5.4 and 5.5 by

Group	indicator of sub-groups found within each cluster
Number of clients	number of clients contained in each sub-group
μ_c	the mean of the rate of extraction c for each sub-group
SEM_c	the standard error of mean for each sub-group

the univariate Gaussian distribution component fitted for the sub-group is represented by $N(\mu_c, \sigma_c^2)$. The standard error SEM_c is $\sigma_c/\sqrt{n}$. n is the number of clients within the sub-group.

Figure 5.3 is an example of cluster 1 on slot 1 on weekday scenario. There are 4 Gaussian sub-groups for this cluster over 61 clients.

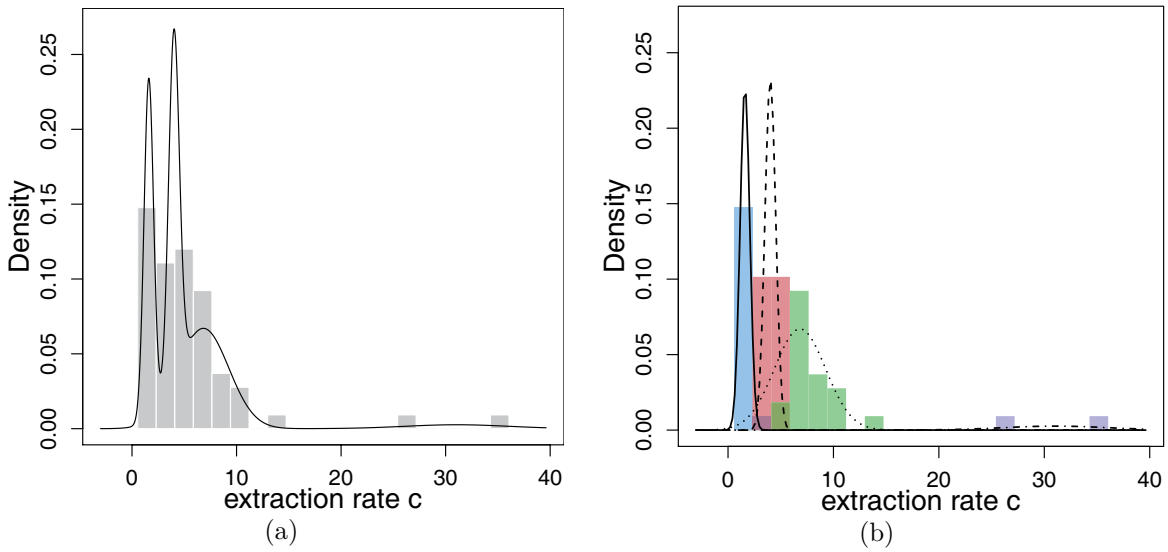


Figure 5.3 Sub clustering results of cluster 1 of slot 1 on weekday. (a) histogram with mixture-based density estimate curve. (b) histograms by sub-group according to the extraction rate c (L/min) with estimated mixture-component densities.

Since the clients within cluster are further clustered into sub-groups according to c, the group size certainly decrease. Some of the groups have only 1 or 2 clients inside, and its Gaussian distribution seems to be less meaningful. Here the model is assumed to be mixed Gaussian, considering the extraction rate c is a constant that describes the demand level, they can be subdivided brutally (e.g. low level less than 5L/min, medium level between 5 and 10L/min, high level greater than 10L/min).

Table 5.3 Clustering results of weekday scenario.

Slot.no	Clustering on λ_0, λ_1							Sub clustering on extraction rate c			
	Cluster	Number of clients	μ_{λ_0} $\times 10^{-4}$	SEM_{λ_0} $\times 10^{-4}$	μ_{λ_1} $\times 10^{-2}$	SEM_{λ_1} $\times 10^{-2}$	ρ	Group	Number of clients	μ_c $\times 10^{-1}$	SEM_c $\times 10^{-1}$
1	1	61	15.7	2.1	46.8	3.4	0	1	17	16.0	1.1
								2	22	40.1	1.1
								3	20	68.3	5.4
								4	2	310.8	35.3
	Outliers	12	-	-	-	-	-	-	-	-	-
2	1	62	166.0	14.0	35.1	2.0	0	1	59	50.1	2.7
								2	3	204.4	38.1
	2	8	289.3	39.1	85.0	5.6	0	1	3	25.5	0.1
								2	3	67.1	7.1
								3	2	237.5	7.7
	Outliers	3	-	-	-	-	-	-	-	-	-
3	1	65	76.9	5.2	36.3	1.8	0.41	1	63	55.9	2.6
								2	2	176.2	14.4
	Outliers	8	-	-	-	-	-	-	-	-	-
4	1	71	77.2	4.4	35.7	1.7	0	1	57	46.2	1.7
								2	14	84.0	9.9
	Outliers	2	-	-	-	-	-	-	-	-	-
5	1	70	125.6	7.5	33.6	1.5	0	1	56	52.0	2.0
								2	14	105.7	10.3
	Outliers	3	-	-	-	-	-	-	-	-	-
6	1	41	36.7	3.5	49.7	5.6	0.11	1	28	42.1	3.8
								2	13	115.0	15.0
	2	29	93.6	10.4	35.8	2.6	0.02	1	29	51.4	3.8
	Outliers	3	-	-	-	-	-	-	-	-	-

Table 5.4 Clustering results of weekend scenario (part 1).

Slot.no	Clustering on λ_0, λ_1							Sub clustering on extraction rate c			
	Cluster	Number of clients	μ_{λ_0} $\times 10^{-4}$	SEM_{λ_0} $\times 10^{-4}$	μ_{λ_1} $\times 10^{-2}$	SEM_{λ_1} $\times 10^{-2}$	ρ	Group	Number of clients	μ_c $\times 10^{-1}$	SEM_c $\times 10^{-1}$
1	1	57	14.0	1.5	36.2	2.6	0	1	53	35.9	2.7
								2	4	136.7	9.9
	2	8	22.2	4.1	127.0	6.9	0	1	2	15.4	0.2
								2	2	39.4	0.2
								3	2	50.5	0.2
								4	1	71.0	0.3
								5	1	224.3	0.3
	Outliers	8	-	-	-	-	-	-	-	-	-
2	1	11	220.9	31.8	43.7	2.0	0	1	1	6.2	0.8
								2	1	24.9	0.8
								3	3	29.8	0.5
								4	1	50.2	0.8
								5	1	57.4	0.8
								6	2	62.1	0.6
								7	2	68.9	0.6
	2	13	112.1	9.9	76.6	8.9	0	1	3	27.9	6.3
								2	4	71.7	1.2
								3	4	94.6	1.4
								4	2	209.6	27.5
		49	85.4	5.3	28.9	1.2	0	1	47	54.3	2.4
								2	2	120.4	11.5

Table 5.5 Clustering results of weekend scenario (part 2).

Slot.no	Clustering on λ_0, λ_1							Sub clustering on extraction rate c			
	Cluster	Number of clients	μ_{λ_0} $\times 10^{-4}$	SEM_{λ_0} $\times 10^{-4}$	μ_{λ_1} $\times 10^{-2}$	SEM_{λ_1} $\times 10^{-2}$	ρ	Group	Number of clients	μ_c $\times 10^{-1}$	SEM_c $\times 10^{-1}$
3	1	68	171.2	9.2	33.4	1.5	0.44	1	68	56.5	2.6
	Outliers	5	-	-	-	-	-	-	-	-	-
4	1	21	168.6	13	47.9	2.8	-0.53	1	21	68.2	7.3
	2	52	97.2	5.2	28.4	1.2	0.41	1	46	49.6	2.2
								2	6	98.8	6
5	1	61	102.2	4.5	33.5	1.6	-0.10	1	58	59.8	3.2
								2	3	176.8	14
	2	9	191.1	25.9	30.9	1.9	-0.02	1	2	29.7	0.2
								2	3	36.9	1.4
								3	2	57.8	0.9
								4	2	83.7	5.3
	Outliers	3	-	-	-	-	-	-	-	-	-
6	1	58	71	5.8	34	1.8	0	1	51	42.2	3.2
								2	7	138.2	8.5
	Outliers	15	-	-	-	-	-	-	-	-	-

CHAPTER 6 CONCLUSION

This chapter firstly makes an overall conclusion and then provides separate conclusions for each section. Furthermore, limitations and recommendations for future study will be offered in this chapter.

6.1 Summary

École Polytechnique's smartDESC project (Natural Resources Canada) has been centered on the use of energy storage (hot water, heated or cooled spaces) naturally present in a power system grid, at customer sites, as a form of distributed battery which could help store excessive intermittent renewable generation, as well as make up for insufficient such generation by reducing the associated loads. In particular, electric water heaters, the storage type that has been of particular interest in this thesis, represent a relatively ubiquitous component whose load could be made to fluctuate so as to help reduce the instantaneous mismatches between power generation and power consumption in an electrical grid. A related major challenge is to be able to do so without compromising the safety and comfort of customers, i.e. essentially ensuring that the load fluctuations remain transparent to them. For this reason, a key step is adequate modeling of the stochastic, time inhomogeneous, customer specific hot water extraction processes, and their segregation at various time intervals during the day, into piecewise stationary, relatively homogeneous customer classes for control purposes. This thesis has proposed approaches towards that goal, when starting from a sufficient sample of electric water heater power consumption data at a collection of customer sites. They rely on a combination of statistical theory, machine learning and stochastic modeling and analysis tools to address (i) time of the day segmentation to address the non stationarity of the statistics of hot water extraction, (ii) estimation of the statistical parameters of piecewise stationary two state Markov chain models of hot water extraction over the corresponding time segments, (iii) clustering of customers into relatively homogeneous water consumption classes according to the time segment of interest during the day.

In Chapter 3, we have presented the tools to distinguish the time segments over 24-hour period in weekdays, and in weekends. These two different time periods were necessary, for a majority of customers, to make their consumption process homogeneous (see Table 3.3). Our approach relied on the so-called fused lasso segmentation method, while the choice of parameters in the lasso technique has been made to be process variance dependent, and automated.

Subsequently, in Chapter 4, we developed a moment based method for the estimation of the two-state time homogeneous three parameter Markov chain $(\lambda_0, \lambda_1, c)$ assumed to represent hot water demand processes over time segments defined earlier. These estimations have been carried out for all customers.

In Chapter 5, a clustering approach is applied to customers in each time segment. The result obtained sub-populations of customers which could be aggregated and controlled collectively within load management scheme distinct time segments.

6.2 Limitations and future prospects

First, only the dataset of the period from November to April is provided in this research. This has limited us to producing results only for the Quebec winter season. However, if we were to continue gathering data during summer time, the methodology developed in this research would still work. Furthermore, one could envision that future EWH's could be equipped with microprocessors and sensors which could record energy consumption data over time. Our estimation algorithms would be implemented to help locally estimate the time varying parameters of the hot water extraction Markov chain. Such parameter sets would be updated over infrequent periods of time (once a month for example) and communicated to the central coordination site via the now ubiquitous smart meters connected to homes. In this manner, the coordinator would be able to maintain an up to date global picture of the dynamics of the loads participating in the load management scheme.

Second, climate conditions affect people's hot water consumption behaviour. Besides the sharp temperature difference between summer and winter, the variations in daily weather could also affect people's hot water temperature preference and water event durations. Weather and ambient temperature, which relate not only to date but also household address, were not provided in the anonymized dataset. If available, a climatic variable parameter could be added to the model to help refine hot water consumption estimation.

Third, the type of activity underlying the water extraction is not specified. Indeed, the file provided contains experimental data of hot water consumption metered in several houses without specifying the particular activity driving that consumption. Ideally, if the activity labels for different hot water usage events could be obtained through a detailed survey or device monitoring, the Markov chain model could be defined as multi-state and have better prediction capabilities, since one would be to able to attach distinct states to distinct activity types and corresponding event duration statistics.

We assumed that the statistics of hot water extraction processes during the time segments

of Chapter 3 are time homogeneous, and the moment based estimation method of Chapter 4 was developed under such assumption. However, the assumption of time homogeneity over the chosen time intervals does appear occasionally questionable although it holds truer on these short intervals as when compared with 24 hours periods (see Figure 4.3 and Figure 4.4). Nevertheless, in Figure 4.2, it is shown that the mean, variance and autocovariance at lag 1 estimated from the actual data record are very similar to those obtained from a simulation using our estimated piecewise homogeneous Markov chain models. In that sense, the modeling results are validated.

REFERENCES

A. M. Abdallah and D. E. Rosenberg, "Heterogeneous residential water and energy linkages and implications for conservation and management", *Journal of Water Resources Planning and Management*, vol. 140, no. 3, pp. 288–297, 2012.

H. Akaike, "Statistical predictor identification", *Annals of the Institute of Statistical Mathematics*, vol. 22, no. 1, pp. 203–217, 1970.

C. Alvarez, R. P. Malhamé, and A. Gabaldon, "A class of models for load management application and evaluation revisited", *IEEE Transactions on Power Systems*, vol. 7, no. 4, pp. 1435–1443, 1992.

T. B. Arnold and R. Tibshirani, "Introduction to the genlasso package".

ATUS. (2003-2011) American time use survey. [Online]. Available: `https://www.bls.gov/tus/#data`

M. Aydinalp, V. I. Ugursal, and A. S. Fung, "Modeling of the space and domestic hot-water heating energy-consumption in the residential sector using neural networks", *Applied Energy*, vol. 79, no. 2, pp. 159–178, 2004.

J. D. Banfield and A. E. Raftery, "Model-based Gaussian and non-Gaussian clustering", *Biometrics*, pp. 803–821, 1993.

J. Bentzen and T. Engsted, "A revival of the autoregressive distributed lag model in estimating energy demand relationships", *Energy*, vol. 26, no. 1, pp. 45–55, 2001.

G. E. Box, G. M. Jenkins, G. C. Reinsel, and G. M. Ljung, *Time Series Analysis: Forecasting and Control*. John Wiley & Sons, 2015.

A. Capasso, W. Grattieri, R. Lamedica, and A. Prudenzi, "A bottom-up approach to residential load modeling", *IEEE Transactions on Power Systems*, vol. 9, no. 2, pp. 957–964, 1994.

G. Celeux and G. Govaert, "Gaussian parsimonious clustering models", *Pattern Recognition*, vol. 28, no. 5, pp. 781–793, 1995.

C. Chong and A. Debs, "Statistical synthesis of power system functional load models", in *Decision and Control including the Symposium on Adaptive Processes, 1979 18th IEEE Conference on*, vol. 18. IEEE, 1979, pp. 264–269.

A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm", *Journal of the royal statistical society. Series B (methodological)*, pp. 1–38, 1977.

P. A. Devijver and J. Kittler, *Pattern recognition: A statistical approach*. Prentice hall, 1982.

S. El-Férik and R. P. Malhamé, "Identification of alternating renewal electric load models from energy measurements", *IEEE transactions on Automatic Control*, vol. 39, no. 6, pp. 1184–1196, 1994.

C. Fraley and A. E. Raftery, "How many clusters? Which clustering method? Answers via model-based cluster analysis", *The Computer Journal*, vol. 41, no. 8, pp. 578–588, 1998.

J. Friedman, T. Hastie, and R. Tibshirani, *The Elements of Statistical Learning*. Springer series in statistics Springer, Berlin, 2001, vol. 1.

A. Fung, M. Aydinalp, V. Ugursal, H. Farahbakhsh, N. Halifax, A. Fung, and D. Ismet, "A residential end-use energy consumption model for Canada", *Canadian Residential Energy End-use. Data and Analysis Center*, 2001.

S. Geisser, *Predictive inference*. CRC press, 1993, vol. 55.

D. N. Gujarati, *Basic Econometrics*. Tata McGraw-Hill Education, 2009.

Y. J. Huang and J. Brodrick, "A bottom-up engineering estimate of the aggregate heating and cooling loads of the entire us building stock", *Lawrence Berkeley National Laboratory*, 2000.

B. J. Johnson, M. R. Starke, O. A. Abdelaziz, R. K. Jackson, and L. M. Tolbert, "A method for modeling household occupant behavior to simulate residential energy consumption", in *Innovative Smart Grid Technologies Conference (ISGT), 2014 IEEE PES*. IEEE, 2014, pp. 1–5.

R. Kadian, R. Dahiya, and H. Garg, “Energy-related emissions and mitigation opportunities from the household sector in delhi”, *Energy Policy*, vol. 35, no. 12, pp. 6195–6211, 2007.

W. Kempton, “Residential hot water: a behaviorally-driven system”, *Energy*, vol. 13, no. 1, pp. 107–114, 1988.

R. Kohavi, “A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection”, in *INTERNATIONAL JOINT CONFERENCE ON ARTIFICIAL INTELLIGENCE*, 1995, pp. 1137–1143.

L. D. Kudryavtsev, *Fubini Theorem*. Encyclopedia of Mathematics, Springer, 2001. [Online]. Available: `http://www.encyclopediaofmath.org/index.php?title=F/f041870`

J.-C. Laurent, G. Desaulniers, R. P. Malhamé, and F. Soumis, “A column generation method for optimal load management via control of electric water heaters”, *IEEE Transactions on Power Systems*, vol. 10, no. 3, pp. 1389–1400, 1995.

J. Laurent and R. Malhamé, “A physically-based computer model of aggregate electric water heating loads”, *IEEE Transactions on Power Systems*, vol. 9, no. 3, pp. 1209–1217, 1994.

M. Lefebvre, *Processus Stochastiques Appliqués*. Presses inter Polytechnique, 2005.

R. Malhamé, “A jump-driven Markovian electric load model”, *Advances in Applied Probability*, vol. 22, no. 03, pp. 564–586, 1990.

R. Mortensen, “Alternating renewal process models for electric power system loads”, *IEEE Transactions on Automatic Control*, vol. 35, no. 11, pp. 1245–1249, 1990.

Natural Resources Canada. Managing Energy Storage Capacities Dispersed in an Electrical Grid to Reduce the Effects of Renewable Energy Source Variability. [Online]. Available: `http://www.nrcan.gc.ca/energy/funding/current-funding-programs/eii/16102`

M. Parti and C. Parti, “The total and appliance-specific conditional demand for electricity in the household sector”, *The Bell Journal of Economics*, pp. 309–321, 1980.

G. Schwarz, “Estimating the dimension of a model”, *Ann. Statist.*, vol. 6, no. 2, pp. 461–464, 03 1978. DOI: `10.1214/aos/1176344136`. [Online]. Available: `http://dx.doi.org/10.1214/aos/1176344136`

L. Scrucca, M. Fop, T. B. Murphy, and A. E. Raftery, "mclust 5: Clustering, Classification and Density Estimation Using Gaussian Finite Mixture Models", *The R Journal*, vol. 8, no. 1, p. 289, 2016.

J. Shao, "An asymptotic theory for linear model selection", *Statistica Sinica*, pp. 221–242, 1997.

F. Sirois, B. Bourdel, and R. Malhamé. (2017, Jul.) Project RENE-034: Managing Energy Storage Capacities Dispersed in an Electrical Grid to Reduce the Effects of Renewable Energy Source Variability – PUBLIC REPORT. [Online]. Available: `http://www.professeurs.polymtl.ca/f.sirois/smartDESC_public_report.pdf`

L. G. Swan and V. I. Ugursal, "Modeling of end-use energy consumption in the residential sector: A review of modeling techniques", *Renewable and Sustainable Energy Reviews*, vol. 13, no. 8, pp. 1819–1835, 2009.

R. J. Tibshirani, J. Taylor *et al.*, "The solution path of the generalized lasso", *The Annals of Statistics*, vol. 39, no. 3, pp. 1335–1371, 2011.

C. J. Wu, "On the convergence properties of the EM algorithm", *The Annals of statistics*, pp. 95–103, 1983.

J. Yang, H. Rivard, and R. Zmeureanu, "Building energy prediction with adaptive artificial neural networks", in *Ninth International IBPSA Conference, Montréal, Canada, August*, 2005, pp. 15–18.

Q. Zhang, "Residential energy consumption in China and its comparison with Japan, Canada, and USA", *Energy and Buildings*, vol. 36, no. 12, pp. 1217–1225, 2004.