TAMPERE UNIVERSITY OF TECHNOLOGY

Chen Tao

Customer Behavior Change Detection Based on AMR Measurements

Master of Science Thesis

Examiners: Prof. Pertti Järventausta and Prof. Hannu Koivisto
Examiners and topic are approved by the Faculty Council of the Faculty of Computing and Electrical Engineering on 5th May 2014.

ABSTRACT

TAMPERE UNIVERSITY OF TECHNOLOGY
Degree Programme in Electrical Engineering
Chen Tao: Customer Behavior Change Detection Based on AMR Measurements
Master of Science Thesis, 60 pages, 1 Appendix page
October 2014
Major subject: Smart Grids
Examiners: Prof. Pertti Järventausta and Prof. Hannu Koivisto
Keywords: Smart Grids, AMR measurement, Customer Behavior, Classification, K-means, Fuzzy C-means

Smart Grids are making use of information and communications technology (ICT) to improve the reliability and flexibility of traditional power grids. This technology depends more and more on methods like load modeling, state estimation, and load forecasting methods. All of these methods are aiming to benefit the whole network analysis to make it become more accurate. Among these methods, load modeling is a very important part of analysis of the whole power network and can offer a good fundamental to other analysis methods. Due to the highly stochastic nature of electricity consumption with many uncertainties, various statistical and classification techniques based on fast data collection are required to help improving the accuracy of conventional load modeling. Nowadays widely used automatic meter reading (AMR) technology in Finland makes it possible to collect customers' hourly load measurements and to use mature clustering methods to analyze those huge sets of customer data and give a better prediction.

In this thesis, some basic classification and regression concepts are borrowed from statistics or machine learning field to help us to analyze the electric customer behavior between different years. This thesis aims to detect either load level change or load shape change of electric customers. K-means and Fuzzy C-means (FCM) are two main methods implemented in MATLAB environment to analyze the load curves. It successfully detects various obvious load pattern changes on different customer types. For the question that when the customer behavior change happens during a year, this thesis can just offer the time information regarding at which week the change happens rather than the specific date. Because we mainly consider the obvious change that lasts for at least one week and ignore temporary changes. The change detection accuracy may be improved in future by more sophisticated methods.

PREFACE

This thesis was done by funding from Smart Grids and Electricity Markets (SGEM) project and examined by Professor Pertti Järventausta and Professor Hannu Koivisto from Tampere University of Technology. I would like to thank Professor Pertti Järventausta for his patient guidance, nice personality and offering me such a precious chance to work on this interesting and valuable topic which matches my interest in data analysis and mathematical modeling. It makes me motivated to learn various statistics and machine learning knowledge, which I enjoyed so much during the learning process. I have found more reasons to strengthen my mathematical background in future. It is a huge regret that at the moment I still can not get the insight of those amazing ideas from those distinguished scholars. But I have confidence that I will learn step by step. This confidence also comes from observing Professor Hannu Koivisto's whole life in academia and his typical scholar office. I would like to thank him for theoretical guidance on every regular meeting and good suggestions for books to read. I think I am not qualified at this moment to claim as a "theoretical guy" as what Prof. Koivisto did. But someday, I will be.

I would also like to thank M.Sc. Antti Mutanen, without whom I almost could not finish this thesis. He guides me in every discussion and gives quite a lot useful advice. And his work is also a huge part of the fundamental of this thesis. Many thanks should also be given to many colleagues in our department with good memory in coffee break, badminton, workshop, conference and leisure time.

Last but not least I would like to thank my parents for their support, care on QQ video every week and understanding my studying abroad far away them.

Chen Tao
5th October 2014

CONTENT

LIST OF ABBREVIATIONS

AMR	Automated Meter Reading
AIC	Akaike Information Criterion
ARMA	Autoregressive Moving Average
ANOVA	Analysis of Variance
BIC	Bayesian Information Criterion
CIS	Customer Information System
DG	Distributed Generation
DR	Demand Response
DNO	Distribution Network Operator
DH	District Heating
DEH	Direct Electric Heating
DSO	Distribution System Operator
FEA	Finnish Electricity Association
FCM	Fuzzy C-means
GSHP	Ground Source Heat Pump
GMM	Gaussian Mixture Model
ICT	Information and Communication Technology
ISODATA	Iterative Self-Organizing Data Analysis Technique
LOM	Loss of Main
NIS	Network Information System
PAR	Periodic Autoregresive
PV	Photovoltaic
SGEM	Smart Grids and Energy Market
SSE	Sum of Squared Errors

LIST OF SYMBOLS

K	the number of clusters
$\mathbf{x}$	input AMR data vector
$\boldsymbol{\mu}$	cluster centre (centroid)
ω_i	notation for ith cluster
$P(\omega_i \mid \mathbf{x}_j)$	probability of data vector $\mathbf{x}_j$ in the ith cluster
d_{ij}	Euclidean distance from data i to data j
N	the number of input data vectors
$P(t)_{TN}$	temperature normalized power consumption at hour t (kW)
$P(t)$	measured power consumption at hour t (kW)
$T_{d,ave}$	daily average of outdoor temperature (°C)
$T_{m,ave}$	long term monthly average of outdoor temperature (°C)
α	temperature dependency parameter (%/°C)
$\mathbf{w}$	weight vector of typical day type
$\mathbf{W}_{diag}$	diagonal matrix whose elements are elements in weight vector $\mathbf{w}$
$\mathbf{E}$	weight vector of yearly consumption
E_j	jth element in vector $\mathbf{E}$

1. INTRODUCTION

Energy crisis and climate change are becoming more and more important issue for all the countries around the world. In recent years, making better use of energy, especially electric energy and improving energy efficiency has motivated us to propose the concept of "Smart Grid", which refers to a class of technology that people are using to bring utility electricity delivery systems into the 21st century, using computer-based remote control and automation [U.S. Dep.Energy 2012]. It enables a number of new functions for power suppliers and power consumers as shown in Fig.1.1.

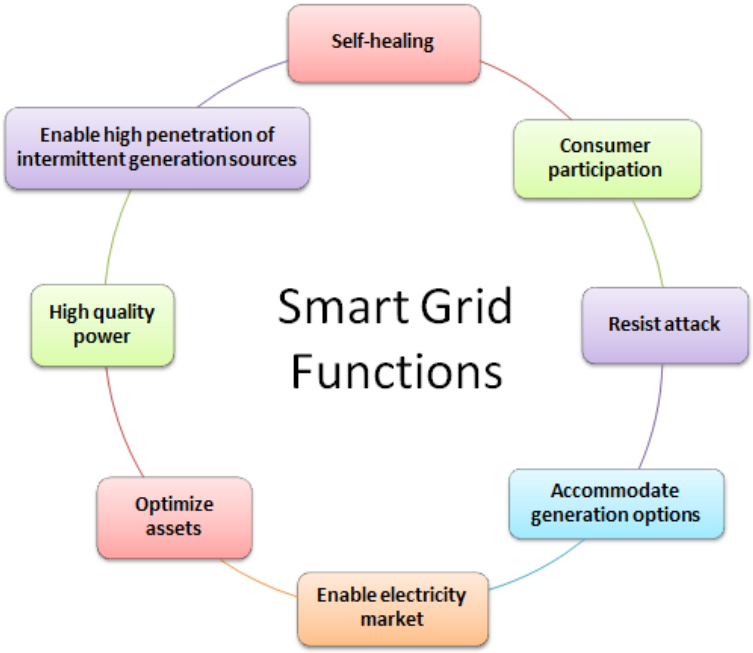


Figure 1.1: Functions of Smart Grids [Smart Grid 2012]

Smart Grids depends heavily on various load modeling, state estimation and load forecasting techniques, which benefit the whole network analysis and make it easier. Among these techniques, load modeling plays a crucial role and offers a good fundamental for other analysis methods. The aim of this thesis is to improve the load modeling accuracy by detecting the customer electricity consumption behavior change and update the customer load modeling information for further analysis.

This thesis is part of the project Smart Grids and Energy Market (SGEM).

In practice, this customer behavior change comes from various resources. For instance, with the advent of smart grids, the ways of operating distribution networks are changing. The integration of distributed generation (DG), like photovoltaic (PV) panel installation, must be considered and detected by the network company due to its bringing bi-directional current flows which may affect fault detection and protecting actions. If some small DG are not detected by the network company in correct way, some problems, like Loss of Main (LOM) protection failure, may arise. Additionally, the idea of demand response (DR) also brings some customer behavior change caused by the response to electricity price or network congestion. Besides, customer behavioral factors identified in [Yao and Steemers 2005] that influence the load profiles are also the main reasons for customer behavior change. They are affected by two root causes: behavioral determinants which are habit driven and relatively flexible; and physical determinants which are driven by environmental factors and building design. Anyway, no matter what is the reason for the behavior change, these behavior changes should be detected promptly in right way. Change detection method will be proposed in this thesis to help improving the load modeling.

In addition, recently more and more network companies emphasize automatic network control and consider more about financial to reduce costs and keep the operation margins low. In such situation, network planning and operation must be made more carefully in order to keep distribution networks within reduced operating margins. In order to achieve this goal, customer class load profiles are widely used in Finland to forecast the short-term or long term load and are helpful in distribution network analysis. It has been shown that load profiles have a big effect on the accuracy of distribution network state estimation [Mutanen et al. 2008]. And when it comes to forecasting the future states of the network, the load profiles have an even bigger role since load prediction depends heavily on load modeling technique and accurate measurements. So if change-detection method can help building more accurate load profiles, it consequently benefits the network analysis. Considering Automatic Meter Reading (AMR) system has already spread quickly in Finland and provided huge amounts of detailed information on customer hourly electricity consumption and behavior information, the change-detection of customer electricity consumption also becomes possible and easier in these days.

But the challenge is that if customer behavior has changed, the data recorded before the change does not represent the customer current behavior any more. Only the post-change data should be used in future load modeling. Thereby, a method for detecting which customers have changed their behavior and when this change happen is needed. Generally, industrial customers have been noted to be predictable as they undertake similar tasks regularly. Domestic customers can vary considerable

from day to day and much of this variability stems from the variability of domestic routine and the appliances in the home. So in most cases, if change happens, it is easier to judge change from industrial customers since they follow the certain load patterns. But for domestic customers, it becomes much harder to judge whether the observed change comes from real load pattern change or just random variation. A standard or limitation must be proposed to separate those real changes and random variation.

In summary, the change-detection method proposed in this thesis is mainly based on classification methods, which borrows some ideas from machine learning field and modify some basic algorithms to apply them for AMR data analysis.

2. BACKGROUND OF CURRENT LOAD MODELING AND CHANGE DETECTION

This chapter introduces some background information about electricity consumption, electricity metering, present load modeling development using AMR measurements in Finland and the necessity for customer behavior change detection. It also includes some work done before in Tampere University of Technology to assume some fundamental to this thesis.

2.1 Electricity consumption in Finland

In Finland, due to the weather reasons, electricity consumption depends heavily on heating system. Heating solution plays an important role in the whole year national electricity consumption as shown in Fig.2.1. At individual household level, heating is also the giant predator accounting over 80% of residential energy consumption as presented in Fig.2.2.

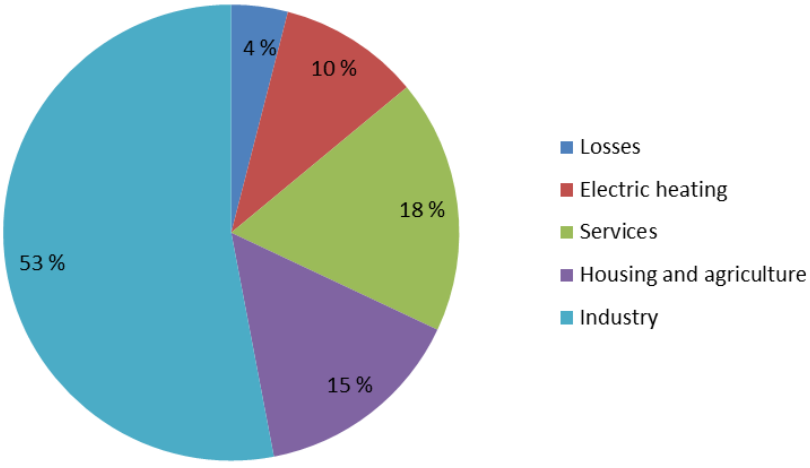


Figure 2.1: Electricity consumption in Finland 2007, 90.3 TWh (www.energia.fi)

Evolving with time, electric customers, especially residential customers, are motivated to transform their heating solution from low energy efficiency to high energy

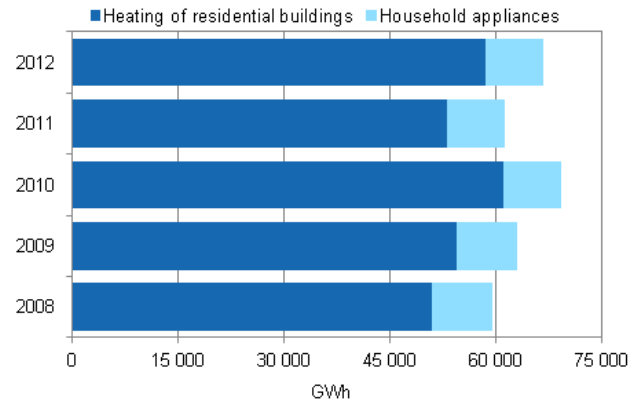


Figure 2.2: Energy consumption in households (Statistics Finland, Published: 13 November 2013)

efficiency due to increasing energy prices. It implies that customers' choosing of heating solution also varies nowadays (see Fig.2.3 and Fig.2.4). In early years, District Heating (DH) was common only in larger cities, but now District Heating already has the largest share of the heating market in Finland. We can observe from Fig.2.3 [Laitinen et al. 2011] that Ground Source Heat Pump (GSHP) also increases its share quite largely due to its direct using of renewable energy and saves electricity consumption compared with commonly used direct electric heating (DEH) [Koreneff 2009]. The heating solution change trend can be even more easily observed from a long term perspective in Fig.2.4 that oil heating decreased a lot between the years 1977-1982.

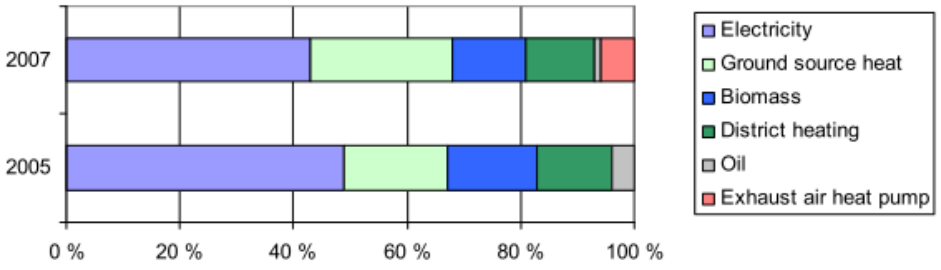


Figure 2.3: Main heating systems of new detached houses in 2005 and 2007 [Laitinen et al. 2011]

Besides, in Finland more than 70% of households own microwave ovens, freezers, dishwashers and so on. Some of them are constantly on, such as freezers and ventilation. Others are used sporadically, like microwave, vacuum and cleaner. This implies those appliances that are used sporadically will cause the large variation in household electricity consumption so that most residential customers have a great potential for load variation. Any appliance use habits change can introduce some change for the

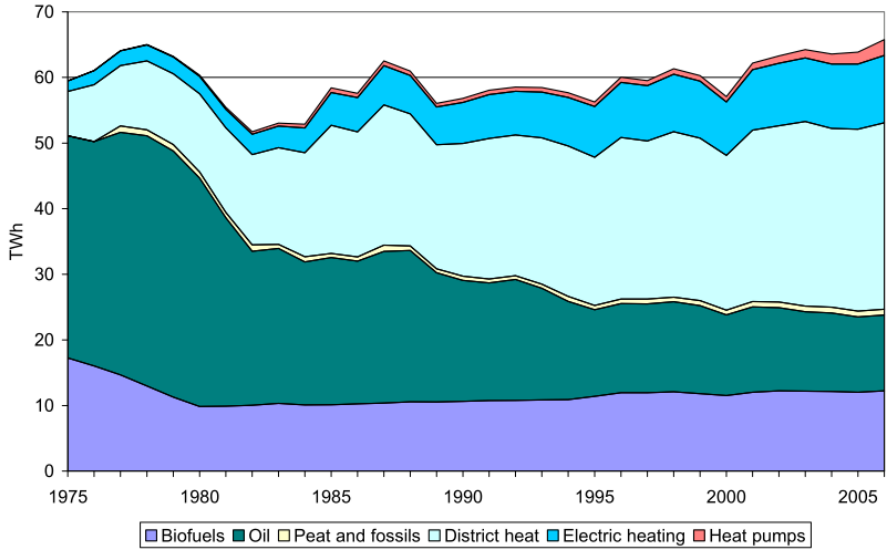


Figure 2.4: Energy sources for heating residential, commercial and public buldings in Finland [Koreneff 2009]

whole day electricity consumption. Additionally, energy saving potential can be observed here according to the results in [Degefa 2010], the savings from installment of heat recovery systems with mechanical supply and exhaust ventilation was very encouraging. The potential of savings after installment of programmable thermostats and predefined settings accounted for about 14.7% of heating consumption. Arranging guidelines and controlling thermostat installations could help significantly improve the energy efficiency of households. It implies there is a great possibility in future that customers will change their electricity consumption in order to achieve the benefit from these savings. And these changes should be detected through some metering methods by Distribution System Operator (DSO) to update Customer Information System (CIS).

2.2 Electricity metering in Finland

Automatic meter reading (AMR) is exactly such a monitoring meter with technology of automatically collecting consumption, diagnostic, and status data from energy metering devices and transferring that data to a central database for billing, troubleshooting, and analyzing. Smart Meter hardware [Darby, 2010] typically comprises a digital replacement to the induction disk type meter, storing readings at a far higher time resolution. It becomes more and more common in many European countries. Nowadays, in Finland, almost all electric customers already have AMR meters. Actually, the installment of AMR meters is encouraged by law in Finland and it requires Finnish DSOs to equip at least 80% of the customers with

AMR meters. So AMR meters provide DSOs with up-to-date electricity consumption data, which can help making load modeling more accurate. AMR meters of this kind can measure quarter-hourly, half-hourly or hourly at each point of electricity consumption. Every measurement value is an aggregation of appliance loads during one certain time interval. However, hourly measurement is the most used in Finland and works as the time resolution for data set in this thesis.

2.3 Load profiling in Finland

Finnish electric utilities started to co-operate in load research in the 1980s. In 1992 Finnish Electricity Association (FEA) published customer class load profiles for 46 different customer classes, 18 of which are for housing and the rest for agriculture, industry and services [Seppälä 1996]. These published load profiles work well for customer load prediction and power network planning purpose. The most prominent shortcoming of these profiles is their age; during the past 20 years electricity consumption has experienced significant changes, the amount of heat pumps and air-conditioners has multiplied, the use of entertainment electronics has increased and electricity consumption in recreational dwellings has changed. Furthermore, in the future, the changes will be even bigger if plug-in hybrids, customer-specific distributed generation and demand response activities become popular. So customer classification and load profiling must be improved with the help of AMR measurements and make sure the above mentioned customer behavior changes can be detected in time for better network analysis.

2.4 Research activities on load modeling at TUT

It is an intuitive idea that different customers can be clustered into similarly behaving groups to make it more convenient for electricity distribution companies to model and predict customer loads. The clustering can be done periodically, for example once a year. If one year time window is used in clustering, changes between years can be taken into account, but the intra-year changes cause errors to customer classification. Moreover, it would be beneficial to use more than one year data in clustering. This customer classification work has already been done in Tampere University of Technology by Antti Mutanen, under Smart Grids and Energy Markets (SGEM) project, which aims to develop international smart grid solutions that can be demonstrated in a real environment utilizing Finnish R&D infrastructure. The customer classification using various clustering methods like K-means, ISODATA, GMM offer good fundamental to further analyze the behavior of customers [Mutanen et al. 2011] [Mutanen 2010] [Stephen&Mutanen et al. 2014]. As Fig.2.5 shows, the Matlab program can read AMR measurements from a database, performs

clustering and exports updated customer classifications and load profiles in to the Network Information System (NIS) [Mutanen 2011]. This has been developed in SGEM project.

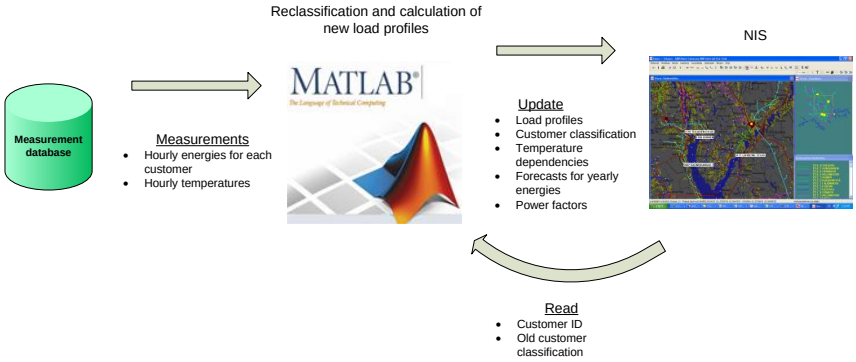


Figure 2.5: Load profiling demonstration [Mutanen 2011]

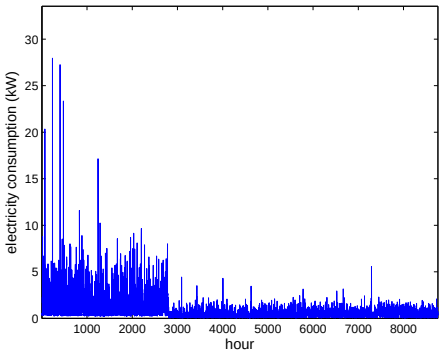
This thesis work belongs also to the same SGEM project and uses the study results of Antti Mutanen. In the report of SGEM project [Mutanen 2010], it is found that load profile updating was more accurate than customer reclassification because the old load profiles did not represent the current electricity consumption habits. It is also possible that the customers did not change their behaviour but the load profiles were incorrect to begin with. Especially for some large individual customers, their behavior changes are important for regional load prediction since most of their load changes are significant if some changes are observed. Thus in order to determine when the load profile updating is needed, we introduce here the customer change detection problem.

2.5 Definition of change in customer behavior

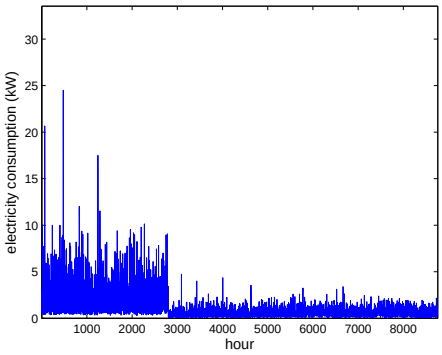
Generally, there are two categories of change defined in different fields. One is defined as intra-year change, namely the measurements or data changing with time series. It is similar to the changes defined in many other fields and used by most literature. Another one is different from the customer behavior point of view, which is defined as between-years change. This means that no change information can be detected at all if we are just given only one year time series data, for example AMR data. We must compare two or more different years AMR data to judge whether some changes indeed happen or not. Because any customer can just repeat his behavior in previous year even some intra-year change happens. For instance, in Fig.2.6b, although some sudden intra-year change can be observed in hour 2800, there is no change compared with year 2009. In contrast, if an artificial consumption figure for this customer appears as shown in Fig.2.6c although there is no obvious intra-year

change, there is between-years change since it changes the pattern compared with year 2009.

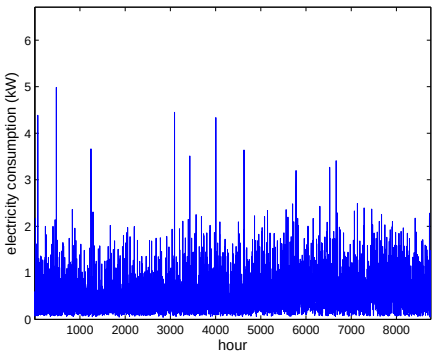
In this thesis, it should be pointed out that "change" here is mainly refered to the between-years change, due to its importance for load modeling and prediction purpose. Thus we detect the changes always based on the comparison between AMR measurements of two or more different years. In this thesis, we further divide this between-years change into load level change and load shape change. It means the comparison will always be implemented either based on their load level or load shape.



(a) AMR measurements for a customer in year 2009



(b) AMR measurements for a customer in year 2010 (no change compared with year 2009)



(c) Artificial AMR measurements for a customer in year 2010 (change compared with year 2009)

Figure 2.6: Intra-year change vs Between-years change

2.6 Motivation for change detection study

The Fig.2.7 shows a typical industrial customer behavior change between 2009 and 2010. This change is probably caused by change in the factory working cycle, such as change from 1-shift to 2-shift caused by socio-economical reason and economic

fluctuation. For some domestic customers, the change may come from various technological and social reasons, such as transfer from oil heating to electric heating, increasing of lighting efficiency, possible connecting of plug-in electric vehicles in future, rearrangement of the holidays, moving out of some tenants.

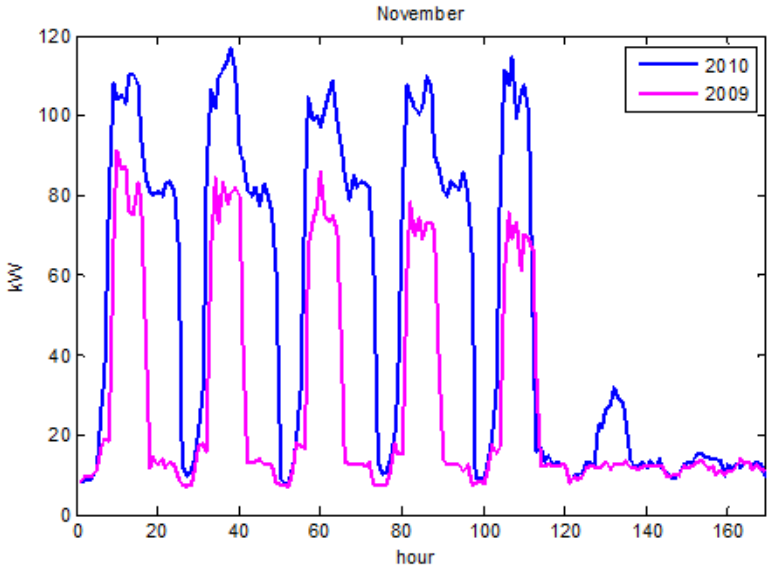


Figure 2.7: The load curve change of a given customer during 168 hours period

If customer behavior changes can be detected correctly and in early time, the accuracy of customer classification and load profiling can be improved significantly by choosing meaningful AMR measurements data after the change point. More accurate load profiles will benefit the network analysis, state estimation, consumption statistics, etc. This customer behavior change information can also be informed to electricity retailers to predict the future power consumption, which may affect the electric market strategy. Besides, with the increasing crucial role that Information and Communication Technology (ICT) solution plays in power industry, change detection can help improving the service from suppliers by offering more accurate customer information and updating more demand information.

3. CHANGE-DETECTION METHODS

This chapter introduces state-of-art change detection methods widely used in various fields. Some of them are difficult to apply in specific load modeling problem due to the different definition of change here and the uniqueness of electricity consumption. But some basic ideas are worth mentioning here to imply their usefulness in further analysis.

3.1 Clustering based method

With the increasing development of data mining, machine learning, artificial intelligent application in power engineering, electric load modeling can utilize various kind of tools, like clustering techniques [Figueiredo et al. 2005]. These clustering techniques can be divided into several categories, such as centroids based clustering, distribution based clustering and density-based clustering. Some clustering methods have already been applied to obtain good enough results, like K-means, Iterative Self-Organizing Data Analysis Technique (ISODATA), and Gaussian Mixture Model (GMM) studied in [Stephen&Mutanen et al. 2014], [Mutanen et al. 2011]. These works offer some fundamental to detecting customer behavior change by analyzing their clustering information. The clustering method K-means, Fuzzy C-means and some issues about choosing a suitable amount of clusters are discussed here.

3.1.1 K-means Algorithm

The aim of the clustering problem is trying to divide a set of objects into different groups such that objects in the same group are more similar to each other than to those in other groups. K-means is exactly such an understandable, easily implemented and widely used clustering algorithm, which divides the input data set into K groups by their similarity. The following introduction comes from some explanations in [Bishop et al. 2006]. Suppose a data set $\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N\}$ consisting of N independent random vectors $\mathbf{x}_j$ with D-dimension. Our goal is to partition the data set into K groups, so called clusters. Before this, the value of K should be chosen as the amount of total clusters. Intuitively, a cluster can be thought as a group, inside which inter-group distances are small compared with the distances to points outside of the group. We can formalize this notion by introducing a set of D-dimensional vectors $\boldsymbol{\mu}_i$, where $i = 1, \ldots, K$, in which $\boldsymbol{\mu}_i$ is a prototype associated with the ith

cluster. Shortly, we can think of the $\boldsymbol{\mu}_i$ as representing the centers of the clusters (centroids). Our goal is then to find an assignment of data points to clusters, as well as a set of vectors $\{\boldsymbol{\mu}_i\}$, such that the sum of the squares of the distances of each data point to its closest vector $\boldsymbol{\mu}_i$ is minimized. In this thesis, $\mathbf{x}_j$ stands for AMR measurement of jth customer and every element in this vector is an hourly electricity consumption value. We can then define an objective function, sometimes called a distortion measure given by

$$J = \sum_{j=1}^{N} \sum_{i=1}^{K} r_{ij} \parallel \mathbf{x}_j - \boldsymbol{\mu}_i \parallel^2 \tag{3.1}$$

$$r_{ij} = \begin{cases} 1 & \text{if } i = \text{argmin}_i \parallel \mathbf{x}_j - \boldsymbol{\mu}_i \parallel^2 \\ 0 & \text{otherwise.} \end{cases} \tag{3.2}$$

which represents the sum of the squares of the distances of each data point to its assigned vector $\boldsymbol{\mu}_i$. K-means algorithm assumes normal distribution of data with the same, shared, diagonal covariance for each cluster. It gives a satisfied clustering result to offer a set of indexes or labels indicating to which cluster or group one multi-dimensional data point belongs. Then further analysis of the change of these indices or labels will indicate the change of the data point. This is the main idea used in this thesis and is able to be improved by Fuzzy C-means to get more informative indices, namely memberships in Fuzzy C-means.

3.1.2 Fuzzy C-means algorithm

The basic idea of Fuzzy C-means is similar to K-means, it just offers some additional information about the probability of a specific point belonging to a certain cluster. So the objects on the boundaries between several classes are not forced to fully belong to one group, but rather are assigned membership degrees between 0 and 1 indicating their partial membership. Thus, the cluster centre is the mean of all data points in the same data set, weighted by their degree of belonging to the cluster [Chen et al. 1996]. In every iteration of the classical K-means procedure, each data point is assumed to be in exactly one cluster, as implied by algorithm 1. But in Fuzzy C-means, we can relax this condition and assume that each sample $\mathbf{x}_j$ has some graded or "fuzzy" cluster membership in every cluster. At root, these "memberships" are equivalent to the probabilities given in following equations [Duda et al. 2000].

$$P(\omega_i \mid \mathbf{x}_j) = \frac{(1/d_{ij})^{1/(b-1)}}{\sum_{r=1}^{K} (1/d_{rj})^{1/(b-1)}}, \quad d_{ij} = \parallel \mathbf{x}_j - \boldsymbol{\mu}_i \parallel^2 . \tag{3.3}$$

where:

$0 \leq P(\omega_i \mid \mathbf{x}_j) \leq 1$

In the resulting fuzzy C-means clustering algorithm we seek a minimum of a global cost function

$$L = \sum_{i=1}^{K} \sum_{j=1}^{N} [P(\omega_i \mid \mathbf{x}_j)]^b \parallel \mathbf{x}_j - \boldsymbol{\mu}_i \parallel^2, \tag{3.4}$$

where $b > 1$ is a free parameter chosen to adjust the "blending" of different clusters. For convenient reason, usually b is chosen to be 2. If b is set to 0, this global cost function is merely a sum of squared errors criterion we shall see again in K-means algorithm. However, the probabilities of cluster membership for each point should be normalized as following to ensure that the sum of all the possibilities belonging to each cluster will be exactly 1.

$$\sum_{i=1}^{K} P(\omega_i \mid \mathbf{x}_j) = 1, \quad j = 1, \ldots, N. \tag{3.5}$$

After simplifying assumptions (see [Duda et al. 2000]), we have the solution for the global cost function (i.e. the minimum of L),

$$\partial L / \partial \boldsymbol{\mu}_i = 0 \quad and \quad \partial L / \partial P_j = 0, \tag{3.6}$$

and these lead to the conditions

$$\boldsymbol{\mu}_i = \frac{\sum_{j=1}^{N} [P(\omega_i \mid \mathbf{x}_j)]^b \mathbf{x}_j}{\sum_{j=1}^{N} [P(\omega_i \mid \mathbf{x}_j)]^b} \tag{3.7}$$

In general, the criterion is minimized when the cluster centers $\boldsymbol{\mu}_i$ are near those points that have high estimated probability of being in cluster i, which is intuitively matched with conventional K-means. Since it is rare to have analytic solutions for such a heavy non-linear equation, the cluster means and point probabilities are estimated iteratively as approximate numerical solutions according to the following algorithm [Duda et al. 2000]:

where:
N is the total amount of customers as input data
K is the number of clusters

Algorithm 1 Fuzzy C-means

begin initialize $N, \boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \ldots, \boldsymbol{\mu}_K, P(\omega_i \mid \mathbf{x}_j), i = 1, \ldots, K; j = 1, \ldots, N$
  normalize probabilities of cluster memberships
  do classify N samples according to nearest $\boldsymbol{\mu}_i$
    recompute $\boldsymbol{\mu}_i$
    recompute $P(\omega_i \mid \mathbf{x}_j)$
  until no change in $\boldsymbol{\mu}_i$ and $P(\omega_i \mid \mathbf{x}_j)$
 return $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \ldots, \boldsymbol{\mu}_K$
end

$\boldsymbol{\mu}_i$ is the centroid of i_{th} cluster

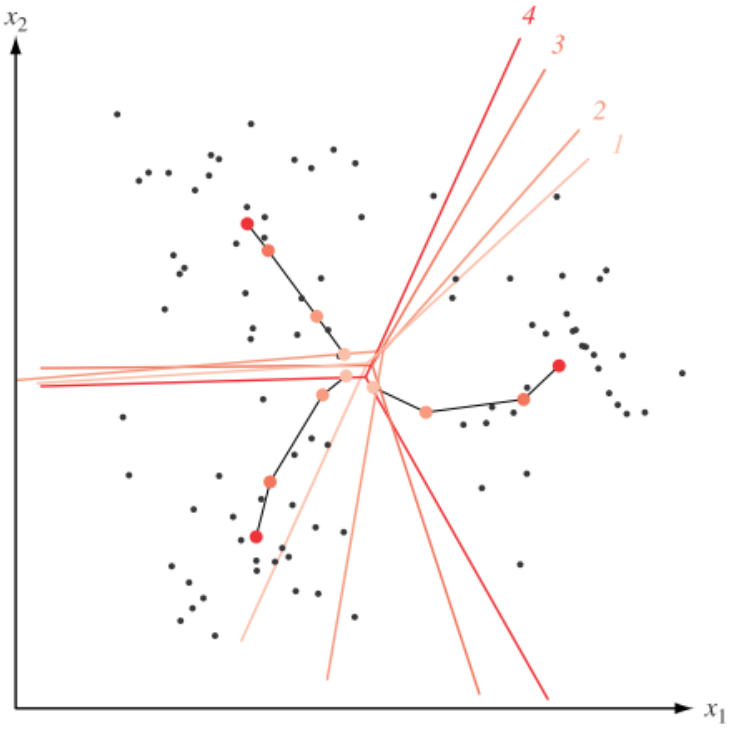


Figure 3.1: Two dimensional data to illustrate the Fuzzy C-means algorithm, red lines are decision boundaries for four iterations (Pattern Classification, R.O.Duda)

The Fig.3.1 illustrates the algorithm for K=3 clusters, at each iteration of the fuzzy C-means clustering algorithm, the probability of category membership for each point are adjusted accordingly. After four iterations, the algorithm has converged. At early iterations the means lie near the center of the full data set because each point has a non-negligible "membership" (i.e. probability) in each cluster. At later iterations the means separate and each membership tends towards the value 1.0 or 0.0. Clearly, the classical K-means algorithm is just of special case where the memberships for all points obey

$$P(\omega_i \mid \mathbf{x}_j) = \begin{cases} 1 & \text{if } \| \mathbf{x}_j - \boldsymbol{\mu}_i \| < \| \mathbf{x}_j - \boldsymbol{\mu}_{i'} \| \text{ for all } i' \neq i \\ 0 & \text{otherwise.} \end{cases} \tag{3.8}$$

While it may seem that such graded membership might improve the convergence of K-means over its classical counterpart, in practice there are also several drawbacks to the fuzzy method. When the number of clusters is specified incorrectly, serious problems may arise.

Fortunately, in this thesis, we just focus on membership characteristics of Fuzzy C-means and mainly borrow the idea of "partially belong to" or "how much degree belong to" one specific cluster. The membership probabilities formula offers us a bunch of membership coefficients of one specific customer AMR data for every centroid.

3.1.3 Number of clusters

Theoretically, choosing the number of clusters can usually be done by observing the "knee-point" of the objective distance or based on certain type of criteria like Bayesian information criterion (BIC), Akaike information criterion (AIC) and Sum of squared errors (SSE). But for our purposes, the choice of the number of customer classes mainly depends on practical aspects, as the willingness of the distribution service provider to create a specific set of tariffs [Chicco et al. 2006]. As such, the number of clusters cannot be too high, in order to allow for easy management of the commercial data. In Finland, Finnish Electricity Association (FEA) published customer class load profiles for 46 different customer classes in 1992, 18 of which are for housing and the rest for agriculture, industry and services [Stephen&Mutanen et al. 2014]. For instance, there is a Distribution Network Operator (DNO) in southern Finland choosing 38 typical electricity consumption groups to decide which group every customer should belong to. And out of those 38 groups, 8 are actually individual customers (classes 31-38), 30 customer group models are left as general models (see Appendix). This chosen amount of clusters is reasonable since it gives enough information about the customer classification and their behavior characteristics. Hence, in further study of this thesis, we choose K=30 as the number of clusters to implement the change detection method based on classification.

3.2 Bayesian method

Detection of change-points is a useful and mature topic already studied in other modeling and prediction of time series areas besides power engineering, such as finance, biometrics, and robotics. So some basic ideas can be borrowed to electric

customer behavior change detection problem to help finding intra-year abrupt variations of electric load.

Among these works, Bayesian framework is a common one and some online methods have been developed and applied in real world [Adams 2007]. Rather than retrospective segmentation, we can focus on causal predictive filtering in electric load modeling; generating an accurate distribution of the next unseen datum in the sequence, given only data already observed [Adams 2007]. By introducing the concept run-length to measure the behavior consistence, we can judge at a specific hour point, whether the load distribution, namely customer electricity consumption behavior, changes or not. This method has been implemented well and shows a meaningful result in a similar way of [Stephen et al. 2014]. It emphasizes on the real-time operation of this load monitoring and mainly analyzes the intra-year customer behavior. This method offers a powerful way to detect such kind of intra-year change detection. Nevertheless, in this thesis we analyze more about the between-years change detection and have a lot of historical data to help analyzing one specific customer. Off-line method is preferred to be used in this thesis and can be combined with some existing clustering methods to give more accuracy comparison. It is possible that in future these online and off-line methods can be combined together to give accurate change-detection results. But in this thesis we mainly focus on the off-line method.

This Bayesian online method is good for some time series change-detection problem but offers little information about the degree of how much the change happens, just informing whether there is a change or not. It is not enough for our case to detect the grade of customer behavior change.

3.3 Regression and Time Series method

Time Series method is applied a lot in econometrics field and deals with highly stochastic stock market problem [Brockwell & Davis 2009]. It is also a common method used in load modeling or load forecasting field, combined with regression and dynamics auto-regression method, which is mainly based on statistical inference and Autoregressive Moving Average (ARMA) model. It mainly runs through predicting short-term or day-ahead load curve to observe the customer behavior in a reasonable range [Wang 2009]. On the other hand, seasonality is an important issue in time series modeling methods. In [Espinoza et al. 2005], a seasonal-modeling approach based on periodic autoregresive (PAR) models is proposed to make short-term load prediction. This PAR model can capture the intra-daily seasonal pattern. Monthly and weekly seasonal are modeled by dummy variables, which is a common method used in electric load forecasting field [Hahn et al. 2009]. In [Espinoza et al. 2005], They also use the stationarity properties of the estimated models to identify typical

daily customer profiles and further to execute some clustering tasks.

For the electric customer behavior change detection problem defined in this thesis, we can determine the change point by comparing two sets of obtained time series model parameters in small time intervals. This method works well for several certain large customers but hard to apply for analyzing various types of customers in long term. And the change time information will be lost if we set the time interval to be too long.

3.4 One-Way Analysis of Variance (ANOVA) method

It is quite natural to propose some statistical methods for this kind of change detection problem. For instance, One-Way Analysis of Variance (ANOVA) can be used to compare the means of two or more groups on one dependent variable to see if the group means are significantly different from each other [Timothy 2011]. In other words, it can be applied to determine whether some difference in two groups are really caused by a change than just by some random variation or different way of sampling. For change detection of customer electricity consumption, the electricity consumption for one specific customer in two different years can be taken as such two individual groups. Then they can be compared using this ANOVA method. But the issue is that this method does not produce any time information and judge the yearly electricity consumption behavior as a whole, which helps little for load modeling purpose.

Although classification method is used as a main method in this thesis, statistical methos can still be combined to further analysis based on classification results.

4. AMR DATA USED IN THIS STUDY

This chapter describes AMR data set from Koillis-Satakunnan Sähkö (KSAT) and Elenia Networks for analysis purpose in the following chapters. Both of these data sets are hourly measured during one or more years, which means every customer has 8760 (i.e. 365 days×24 hours) measured values stored as a column vector. If this vector is plotted, then we can get a figure of load curve or load profile. Correspondingly the changes of those AMR measurements in different years can also be easily observed as the shape changes of load curves. So in the following chapters, we will use the terms load, load curve, load profile, AMR data interchangeably. On the other hand, the recorded temperature, date and other information are also given as supplementing data to help analyzing customer behavior.

4.1 KSAT data

In KSAT data set, each customer is labelled with a customer class number to show to which class one customer belongs (38 classes in total). The meanings of these class numbers are listed in appendix A.

Every customer is measured hourly through the whole year. In the original data set, there are 3584 customers in total and measured through two complete years, 2009 and 2010. However, before analyzing these customer behaviors, some customers with empty consumption should be removed because there is no customer behavior pattern at all in an empty house. After this, 3577 customers are left in this KSAT data set and 7 empty consumption customers are removed to ensure the accuracy of the following analysis. And the customer number discussed following is also ordered without considering these 7 empty customers.

The AMR data set is arranged as a matrix where each row contains consumption over time and each column corresponds to every evaluated customer. Every customer AMR measurement is saved into a multidimensional column vector with every element stands for the electricity consumption (kWh) in one specific single hour. It should be noticed that here value 0 means zero consumption at this specific hour and value NaN (in MATLAB data format) means measurement is missing at this specific hour. The following Fig.4.1 is an example of AMR measurement values for customer No.444 with class label 3. According to the customer class information table, it belongs to the type of "Housing+partial storage electric heating".

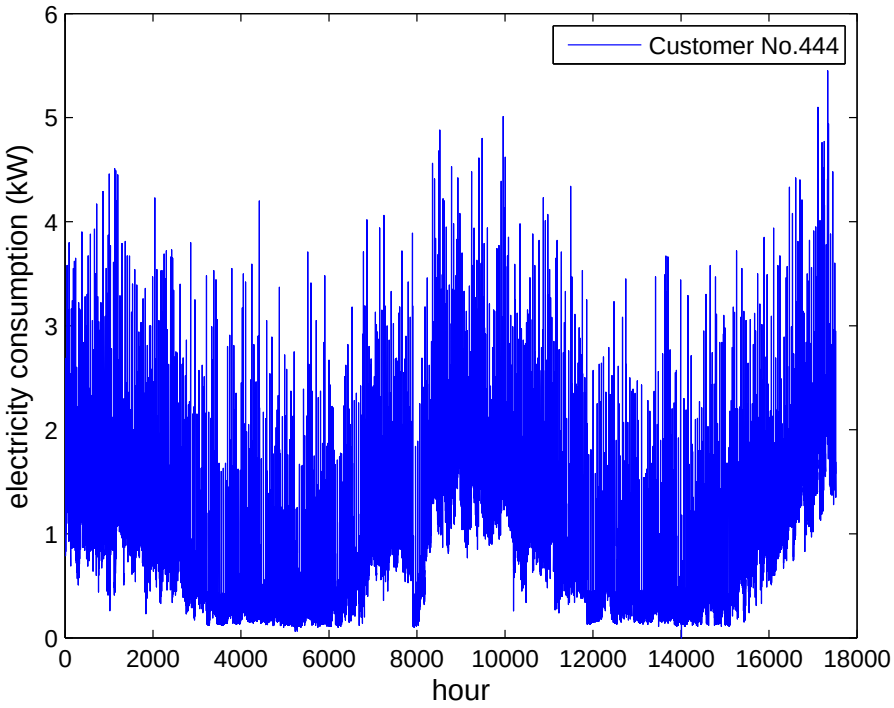


Figure 4.1: Two years (2009-2010) hourly AMR measurement for customer No.444

In Fig.4.1, AMR measurements are recorded in every hour during the whole two years, so in total it has 17520 data points (730 days $\times$ 24 hour/day). The seasonality can also be observed in the figure since every summer around hour 4000-6000 for year 2009 and hour 12000-14000 for year 2010, this customer reaches the consumption valley, which implies it consumes less energy in summer than winter due to temperature effect.

The following Fig.4.2 is another AMR measurement for customer No.666 with label 8 (row house/apartment, no electric heating, no sauna stove) to show some obvious customer electricity consumption behavior change during two different years.

We can observe that there is no seasonality at all in this customer behavior and the consumption behavior changes a lot from hourly 0.2 kW level in the year 2009 (hour 1-8760) to hourly 0.8 kW level in the year 2010 (hour 8761-17520). The load curve is also quite irregular without any certain routine to follow. This phenomenon can perhaps be explained by the fact that the electricity consumption of this customer is not sensitive to the temperature because it has no electric heating and house living situation may be different between the year 2009 and 2010 due to some vacation reason or living elsewhere.

Some more customer cases will be presented in the following chapters to give more typical customer load curve patterns to be analyzed. It should be emphasized again that electric customer behavior is highly stochastic and hard to be compared

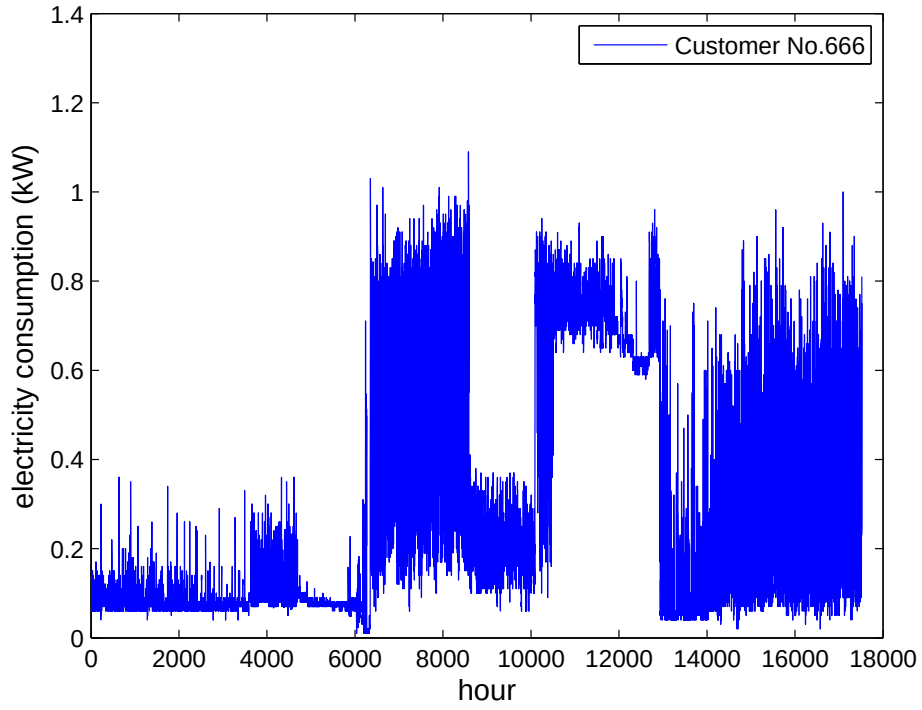


Figure 4.2: Two years (2009-2010) hourly AMR measurement for customer No.666

hour by hour or even day by day. Some more general comparison and classification methods must be used to solve this issue.

4.2 Elenia data

Elenia Oy is an independent distribution system operator servicing over 410000 distribution network customers in approximately 100 municipalities with a network area of nearly 5000 km^2. The data set used here contains the AMR data of 7532 low voltage customers in a small region. This region is a small town with around 10,000 inhabitants. There is a HV/MV (110/20 kV) primary substation feeding all these inhabitants. In this area, there are nine MV feeders, of which two are mainly underground cable and seven are mainly overhead line having total of 457 km of 20 kV MV network. It also has about 469 distribution transformers (20 kV/400 V) and 793 km of 400 V network. An example of customer No.5 in this Elenia data set is shown in Fig.4.3.

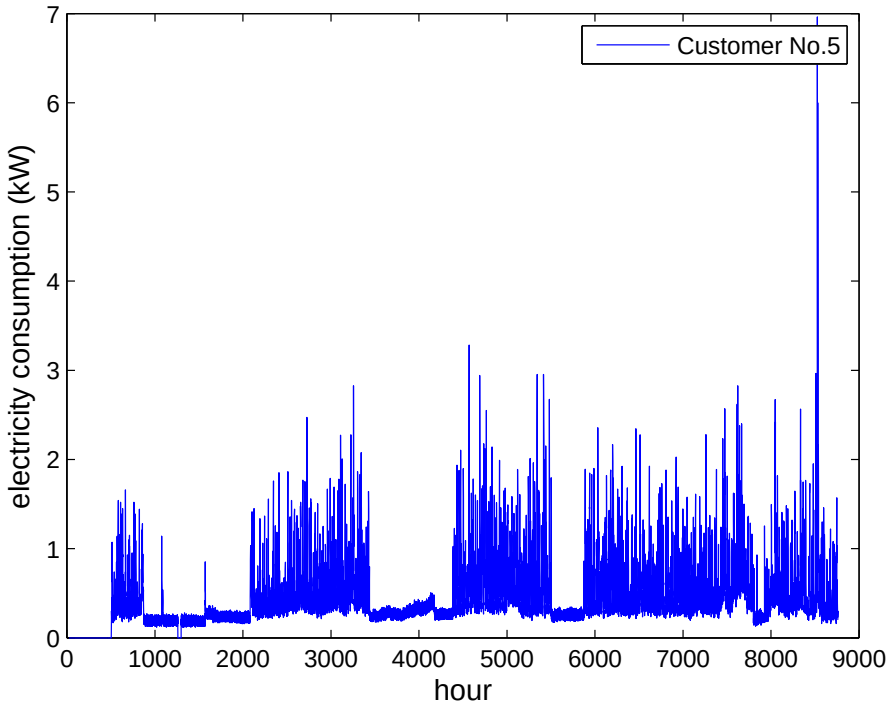


Figure 4.3: Hourly AMR measurement for customer No.5 in Elenia data from June 2010 to June 2011

This Elenia data set is mainly used for test purpose in Chapter 8 but also used as an example data to show the customer reclassification problem in Chapter 6. Originally, it includes both low voltage and medium voltage measurements. Here in this thesis, only the low voltage customer data is used.

5. CHANGE DETECTION OF ELECTRICITY CONSUMPTION LEVEL

This chapter proposes a basic time window method to detect the obvious load level change. Actually, it is hard to separate pure load level change from load pattern change completely. But in general, due to some long lasting low consumption behavior, some huge step changes can still be clearly observed from some customer load curves. Obvious step changes of this kind can be detected very well with this time window.

5.1 Typical load level change

Electric customer load level changes can be caused by various factors from increasing energy use efficiency to travelling on vacation. It is found that most load level change or consumption step change will generally last for some time, a few weeks, several months or even a whole year. So some randomly happened temporal load level change should not be considered seriously. In fact these temporal changes contribute little to the whole year electricity consumption since they just last for a few hours or a few days.

The Fig.5.1 shows a customer with a few months load level change and the Fig.5.2 shows a customer with a whole year load level change.

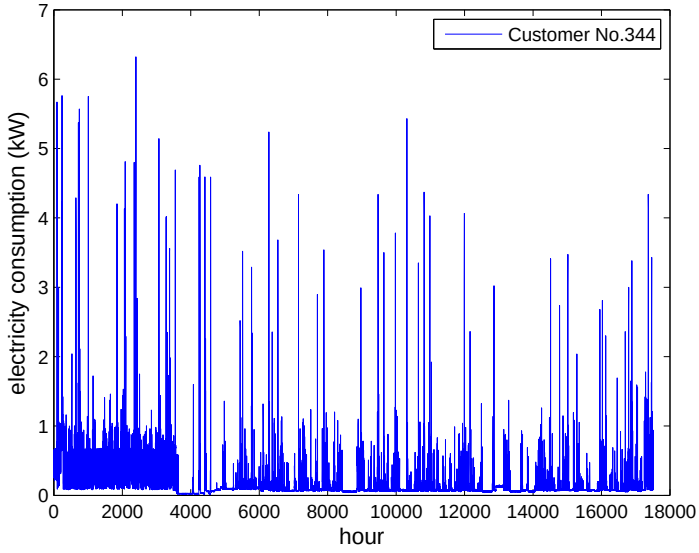


Figure 5.1: Customer No.344 with a few months load level change from about January to April between 2009 and 2010

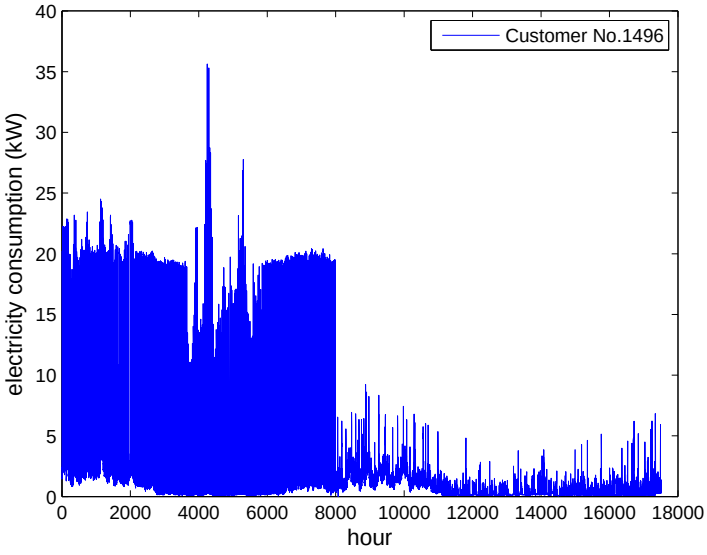


Figure 5.2: Customer No.1496 with whole year load level change, 2009-2010

5.2 Time window method to detect load level change

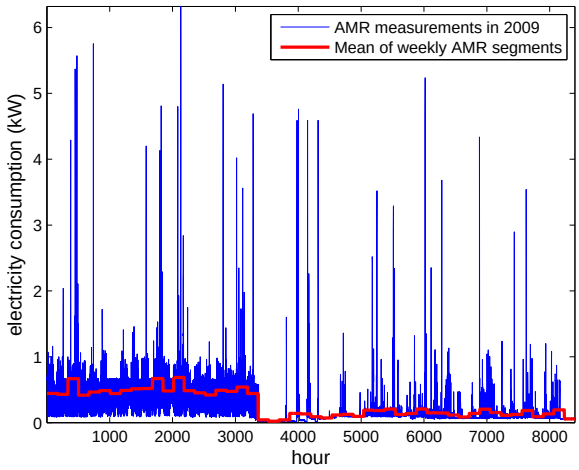
The time window method developed in this thesis tries to divide the whole year AMR data of one specific customer into weekly segments and then compares these weekly segments consumption (load level) respectively in different years. In order to compare load level independent from temperature conditions in different years, all the AMR data are processed by temperature normalization. Since customer behavior

in weekdays and weekends is different, it is unreasonable to directly compare the 1st hour consumption of the year 2009 with the 1st hour consumption of the year 2010. Because it is possible that the 1st hour AMR data of the year 2009 comes from Monday or Tuesday and the 1st hour AMR data of the year 2010 comes from Saturday or Sunday. That is why in following we always divide the whole year AMR data into 50 complete weeks according to the first Monday position of every year. Namely, original 8760 hourly measured AMR data in every compelete year is divided into 50 (weeks) 168 hours dimensional AMR data segments as shown in Fig.5.3. So after this processing the whole year 8760 dimensional AMR data is pruned to 8400 dimensional AMR data. The rest 360 data points are abandoned. This processing may not be so necessary here if we consider the weekly means. But because load shape change detection needs such processing and in order to implement these two kinds of change detection methods for the same dimensional data at the same time, we still do so here.

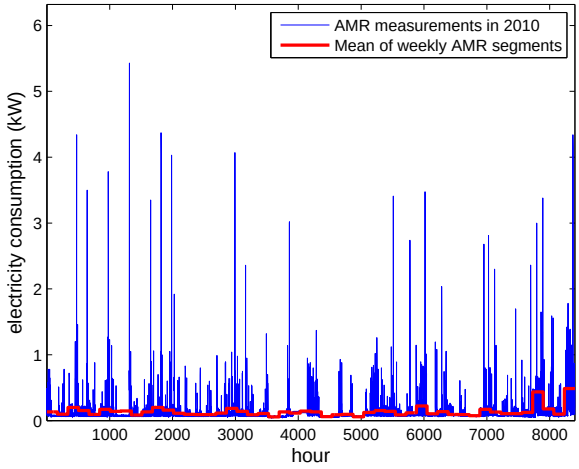
Then we compare the weekly load level in different years based on weekly averages. For instance, the 1st week load level in year 2010 is compared with the $\pm 125\%$ band of 1st week load level in year 2009. The comparing result of whole year 50 weeks can be observed in Fig.5.3. Threshold is set to be that if weekly average consumption in year 2010 exceeds level change detection band of year 2009 at least ten times, then we declare some changes are detected.

It should be pointed out that actually this mean of weekly consumption is calculated by curve fitting function $P = polyfit(X, Y, N)$ in MATLAB with setting $N = 0$. So in fact this mean calculation is linear regression with 0 order polynomial, which follows the Least mean squares (LMS) rule. Besides the reason for convenient figure plotting, we also hope to make this method easy to be modified for various other order polynomial curve fitting for weekly AMR data segments. It brings some idea to also observe the load pattern change by comparing the parameters of weekly AMR data segments curve fitting. But the result of this time window method is not good enough for load pattern change detection, while it works well for load level change detection as shown above.

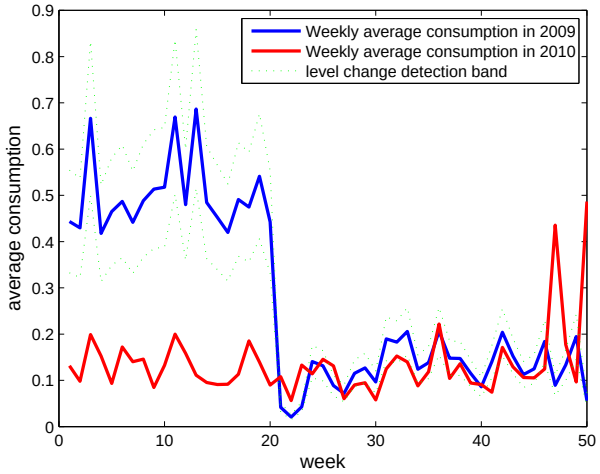
In Fig.5.4, we can observe that the customer No.2455 behaves quite different in two different years. There is obvious load level change in a few beginning months and load shape change in middle year. But the load level in the middle part of two different years is by coincidence so similar to each other, so the threshold band in Fig.5.4c obviously fails to detect such shape change. So in the following chapters, we will introduce several classification methods instead of simply comparing the consumption levels to solve this load shape change detection problem.



(a) Weekly AMR data segments of customer No.344 in year 2009

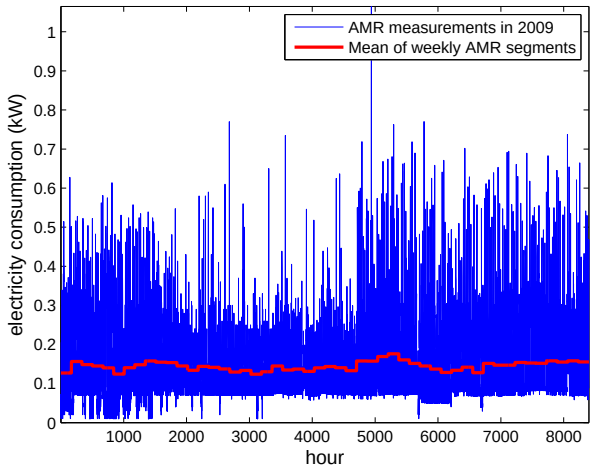


(b) Weekly AMR data segments of customer No.344 in year 2010

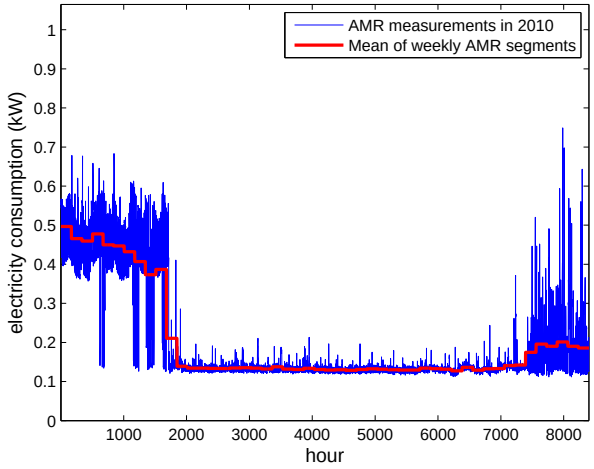


(c) Difference of weekly consumption for customer No.344 between year 2009 and 2010

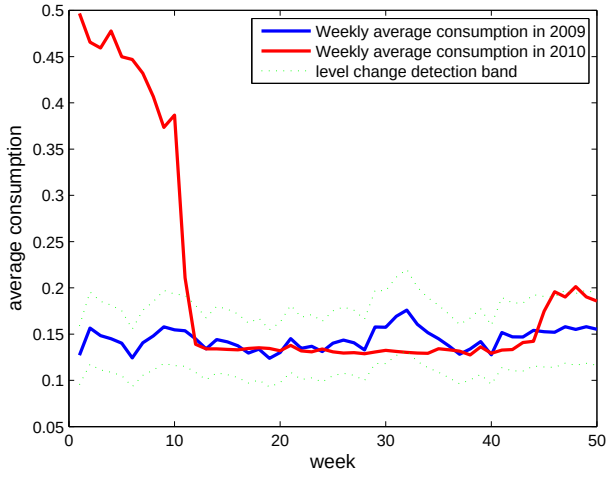
Figure 5.3: Load level change detection for Customer No.344



(a) Weekly AMR data segments of customer No.2455 in year 2009



(b) Weekly AMR data segments of customer No.2455 in year 2010



(c) Difference of weekly consumption for customer No.2455 between year 2009 and 2010

Figure 5.4: Load level change detection for Customer No.2455

6. CLUSTERING METHOD TO DETECT OVERALL BEHAVIOR CHANGE

This chapter introduces the clustering methods for detecting overall electric customer behavior change. The clustering method is different from load level change detection method, which is done through comparison based on weekly AMR data segments. Instead, the clustering method checks the classification of all customers based on annual AMR data to find any clustering result changes happening between different years. This method is quite intuitive and widely used in other fields [Jain et al. 1999].

6.1 Load distribution of electric customers

In order to study electric load modeling and implement our classification algorithm, load distribution must be studied beforehand to offer a theoretical foundation for further analysis. Modeling of the stochastic component of the electrical network load is done in many papers [Stephen&Mutanen et al. 2014][Neimane et al. 2001]. It has been shown that most uncertainties of active and reactive daily peak loads can be modelled by normal distributions [Filho et al. 1991]. Nevertheless, some papers suggest that the best representation of low-voltage network load is beta probability distribution [Herman et al. 1993]. Others [Neimane et al. 2001] use three probability density functions (i.e. normal, log-normal and beta distribution) to model variations of the network load and obtain a conclusion that all three distributions provide a reasonably good representation of load variations. However, if variations of the modelled parameter are non-symmetrical, lognormal or beta distribution would give a better approximation [Meldorf et al. 2007]. In this thesis, through studying customers' AMR measurements, there is no obvious reason to refuse the convenient assumption that normal load distribution is good enough to model the electric customer load curve. For instance, a study of one customer's load distribution is shown in Fig.6.1. Hence, this load distribution assumption offers a solid foundation for various clustering methods in the following sections.

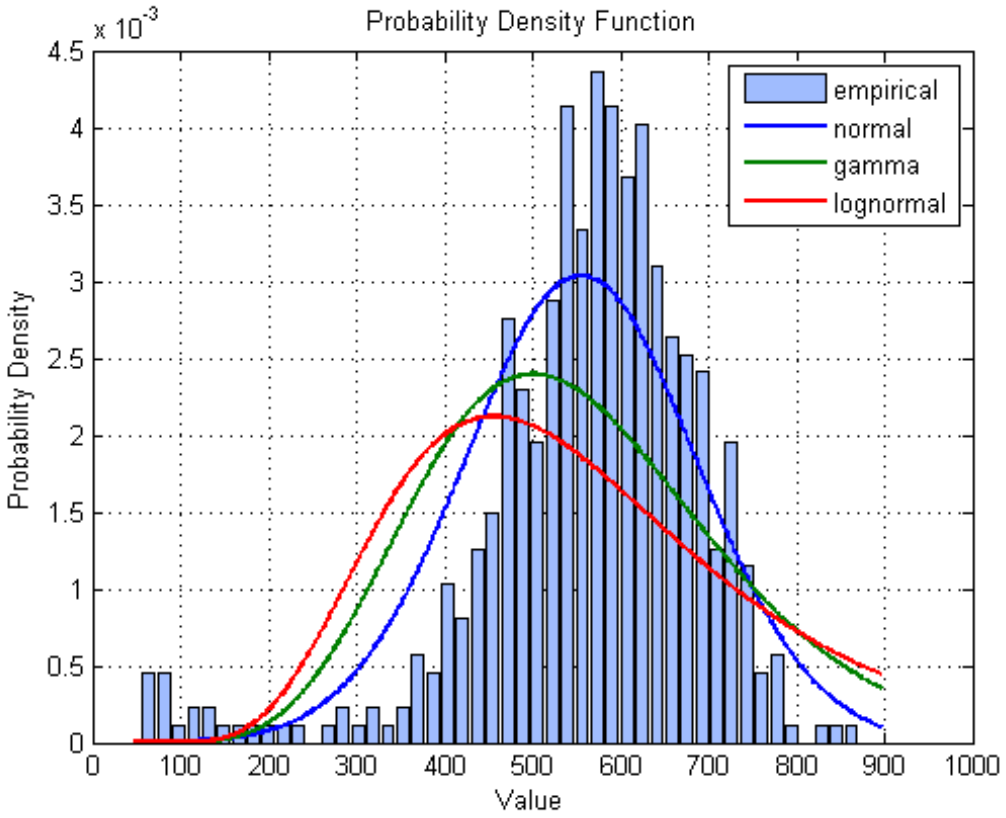


Figure 6.1: Histogram and fitted distribution for hour period 14:00-15:00 in year 2009 and 2010 (weekdays only)

6.2 Customer classification based on AMR measurements

Electricity customer classification can be realized by a pattern-recognition method proposed in [Mutanen 2010]. This method classifies customers into clusters, for which load profiles can be calculated based on AMR measurements. Then these load profiles are used to model customer loads in the distribution system. The method involves temperature dependency correction and outlier filtering [Mutanen 2010]. Due to cold weather in Finland, customer behavior is sensitive to temperature change since electric heating consumption depends heavily on the outdoor temperature as shown in Fig.6.2.

It is obvious that electricity consumption is inversely proportional to outdoor temperature within normal temperature variation range. However, some extreme hot weather may not follow this rule, since high temperature may bring some additional electricity consumption of the air conditioner. Due to the fact that extreme hot weather is a very rare case for Finland, the temperature normalization mainly aims to work within the normal temperature range and sets certain limits for exceptions. Therefore, in order to detect real customer behavior change besides the ones caused by seasonality, some data preprocessing is necessary to remove the effect of

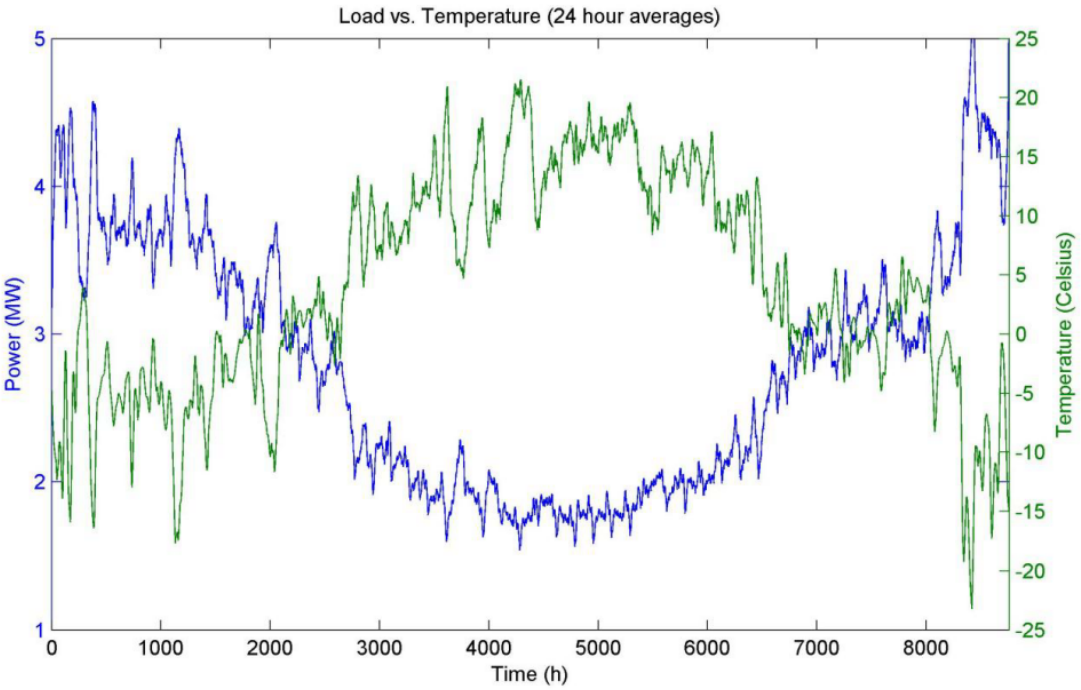


Figure 6.2: Sum of measured loads and temperature for year 2009 (24 hour averages).

temperature. After this, the dimensionality of input data for clustering also needs to be considered, since the raw AMR data is high dimensional up to 8760, which means the load is measured hourly for everyday during the whole year. Thus some dimension reduction techniques can be used to speed up the clustering process, such as the pattern vectorization method used in this thesis. This technique is trying to reduce the high dimensionality of AMR data but without losing customer energy consumption information at the same time. This dimensionality reduction technique can offer satisfied foundation for daily load profiles method discussed in the following section. But later weekly load profiles method needs more complete information on everyday consumption. So direct running of K-means algorithm on this temperature normalized AMR data without dimensionality reduction is proposed in Chapter 7.

6.2.1 Temperature normalization of AMR data

Temperature normalization of AMR measurements can be done in such a way that we assume that temperature sensitive part of the load is linearly dependent on the temperature. In [Mutanen 2010], a linear regression model is proposed to obtain temperature dependence parameter α with confidence level 95%. The temperature dependency parameters are calculated with linear regression analysis for every four seasons separately. The percent error between the daily energy consumption and the average daily consumption on a similar day during that month is chosen as the dependent variable (regressand). And the difference between the daily average temperature and the average temperature on a similar day during that month is chosen as the determining variable (regressor). The significance of relationship between the daily energy and outdoor temperature is assessed with the correlation coefficient and the Student's t-test in [Mutanen 2010]. Hence the temperature normalized load in this thesis can be calculated as following:

$$P(t)_{TN} = \frac{P(t)}{1 + \alpha(T_{d,ave} - T_{m,ave})}, \tag{6.1}$$

where:

$P(t)_{TN}$: is the temperature normalized power consumption at hour t,
$P(t)$: is the measured power consumption at hour t.
$T_{d,ave}$: is the daily average of outdoor temperature,
$T_{m,ave}$: is the long term monthly average of outdoor temperature,
α: is the temperature dependency parameter %/°C.

The temperature normalization is made according to daily average temperatures. The temperature dependency parameter α is assumed to be the same for all hours of the same day. After this temperature normalization process, the temperature normalized AMR data will be obtained to be used as input for K-means clustering algorithm. Such that we can compare electricity consumption in different years with different temperature.

6.2.2 Pattern vectorization of temperature normalized AMR data

Pattern vectorization of temperature normalized AMR data is implemented based on typical day type representation. Here we assign one of three day type (i.e. Weekday, Saturday, Sunday) label to every daily AMR data segment. Then in every

individual month we synthesize all the 30 or 31 daily AMR data segments into just three daily pattern vector segments according to their typical day type labels. It should be noticed that in clustering the daily pattern vector segment with label weekday weighs 5 compared to other two daily pattern vector segments weighing 1 because it includes five days (Monday, Tuesday, Wednesday, Thursday, Friday) consumption information.

Alternatively, we can also synthesize all the 30 or 31 daily AMR data segments into seven daily pattern vector segments for every month according to seven different kinds of day type labels (i.e. Monday, Tuesday, Wednesday, Thursday, Friday, Saturday, Sunday). In this case, every daily pattern vector segment weighs 1 instead of above mentioned weekday pattern vector segment weighing 5. This weight vector will work as a supplement column vector $\mathbf{w}$ to be given to running the K-means algorithm:

$$d_{ij} = \| (\mathbf{x}_i - \boldsymbol{\mu}_j)^T \mathbf{W}_{diag} (\mathbf{x}_i - \boldsymbol{\mu}_j) \|^{1/2} \tag{6.2}$$

where:

d_{ij}: is weighted distance between ith data vector to jth centroid,
$\mathbf{W}_{diag}$: is a diagonal matrix whose elements are elements in weight vector $\mathbf{w}$,
$\mathbf{x}_i$: is ith data vector,
$\boldsymbol{\mu}_j$: is jth centroid.

After temperature normalization and pattern vectorization of the raw AMR data, those 864 (i.e. 12 months×3 typical day type×24 hours) dimensional or 2016 (i.e. 12 months×7 typical day type×24 hours) dimensional pattern vectors are given as input data to K-means algorithm programmed in MATLAB to go through clustering process.

6.2.3 Clustering with weighted K-means algorithm

The weighted K-means algorithm is the modified version of the basic K-means just considering the annual energy level weight for every customer and typical day type weight for every daily time interval in addition. It means that when we calculate the mean (i.e. centroid) of a specific cluster, different customers contribute differently to the mean value according to their yearly energy consumption level. The resultant centroids are always inclined to the high consumption customer, which is a reasonable consideration of the heavy industry (e.g. metal industrial) customer. Because this kind of customer always contributes more to the form of typical indus-

trial electricity consumption group. In MATLAB, this yearly energy level weight is represented by a row vector **E** to weigh the mean calculation of a specific cluster:

$$\boldsymbol{\mu}_i = \sum_{j \in C_i} (\mathbf{x}_j \frac{E_j}{\sum_{j=1}^{N} E_j}) \tag{6.3}$$

where:

$\boldsymbol{\mu}_i$: is ith centroid,
C_i: is a index set representing customer indices grouped in ith cluster,
$\mathbf{x}_j$: is jth customer data,
E_j: is an element storing yearly consumption of jth customer in vector **E**.

The algorithm is like following:

Algorithm 2 Weighted K-means

```
1: begin initialize N, K, μ_1, μ_2, ..., μ_K
2:     do classify N samples according to nearest μ_i
3:       recompute μ_i with weight vectors E and w
4:     until no change in μ_i
5:   return μ_1, μ_2, ..., μ_K
6: end
```

N is the total amount of customers as input data
K is the number of clusters
μ_i is the centroid of i_{th} cluster
E is the weight vector for yearly energy level
w is the weight vector for typical day type

After running the weighted K-means algorithm, the classification result is shown in Fig.6.3. It can be observed that some clusters or groups have much more customers than others, which means in this data set of electric customers, some certain type of customers are dominant. However, it is possible that if a customer changes his behavior obviously in next year, he may move to other clusters (i.e. electric customer group) instead of remaining in the same group.

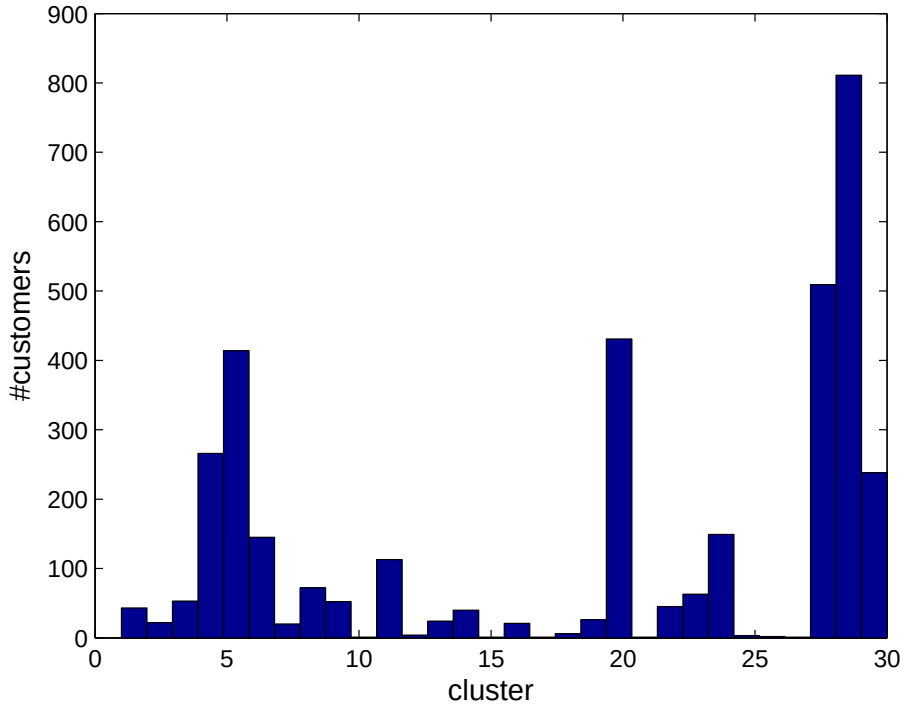


Figure 6.3: Classification result with pattern vectors of AMR data in year 2009

6.3 Reclassification method to detect load shape change

Due to some reason of transforming from direct electric heating to storage electric heating or having some extra electric devices, the customer may be represented by different load curve shapes and consequently be classified to different clusters in different years. This kind of overall year load shape change can be detected easily from the final clustering index of one specific customer. Take customer No.11 from Elenia test data as an example, we classify all the customers into 37 classes (this amount of classes is by default used in Elenia data) and then find that customer No.11 belongs to class 9 in the year 2011. Since the pattern vector of customer's AMR measurements is built by one typical week presentation in every month, it is 2016 dimensional (i.e. 12 months×7 days×24 hours). In the following years 2012 and 2013 it is compared to every other cluster centroid to be reclassified to the nearest class using Euclidean distance. These three years pattern vectors can be observed in Fig.6.4.

According to the reclassification result of every one of these three years, the customer remains in class 9 in the year 2012 and change to class 22 in the year 2013. This can be explained by the obvious pattern vector fluctuation of 2013 compared with other two years. Thus it is a good example to show the reclassification method implemented every year can be helpful to detect the obvious load shape change.

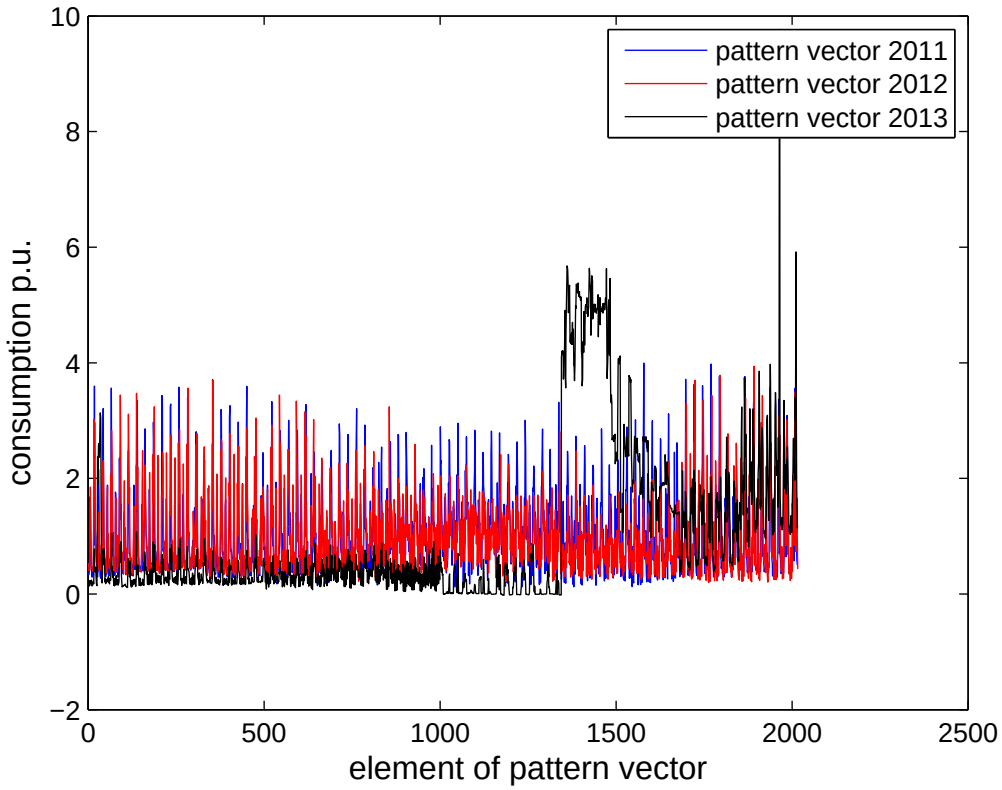


Figure 6.4: Pattern vector of customer No.11 in three years 2011, 2012 and 2013

By studying of all the 7532 customers in Elenia test data during these three years (2011, 2012 and 2013), it can be found that only 2714 customers remain in their original classes in all the three years. About 64% of all the customers have some load pattern change and are reclassified into other classes. So it is fair to say load shape change detection is a common issue for most customers. Some more accurate load shape change detection methods should be proposed to offer further information about the moment of load shape change. For those changes that are not big enough to cause moving of reclassification results (i.e. still remain in the same classification group but indeed change more or less), some informing about the partial change should also be offered. This kind of problem will be discussed more in the following chapter.

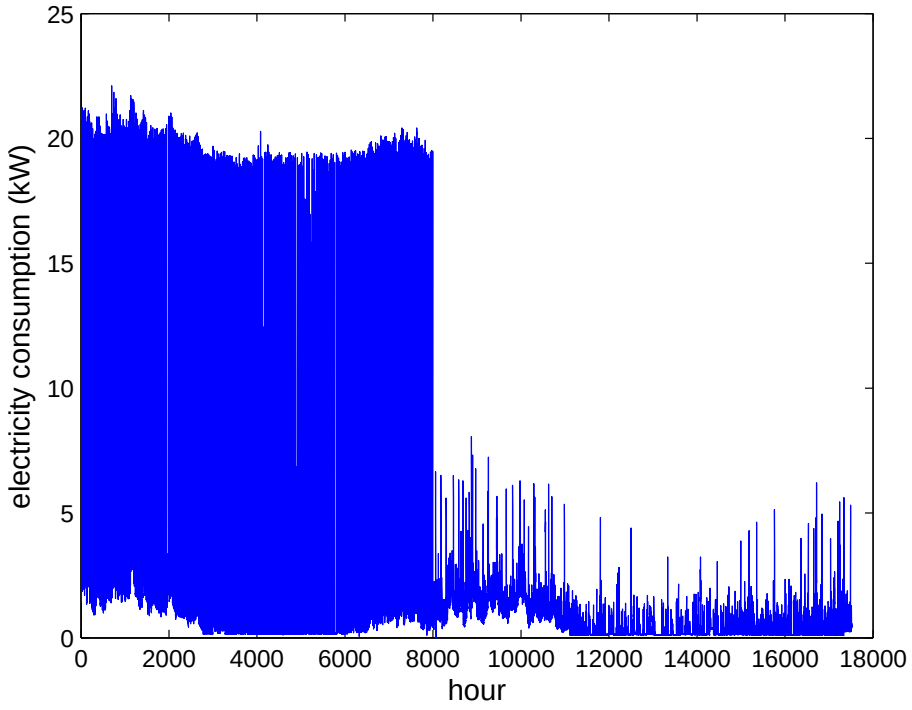
6.4 Daily clustering methods to detect daily load shape change

Similar to electric customer classification based on annual AMR data, we can also classify customer based on their daily AMR data to observe the daily classification results. If at similar day of two different years (e.g. 1st Monday, 2009 and 1st Monday, 2010), the daily load shape change is big enough to affect the daily classification results (e.g. classified to cluster No.8 in 1st Monday, 2009 versus classified to cluster

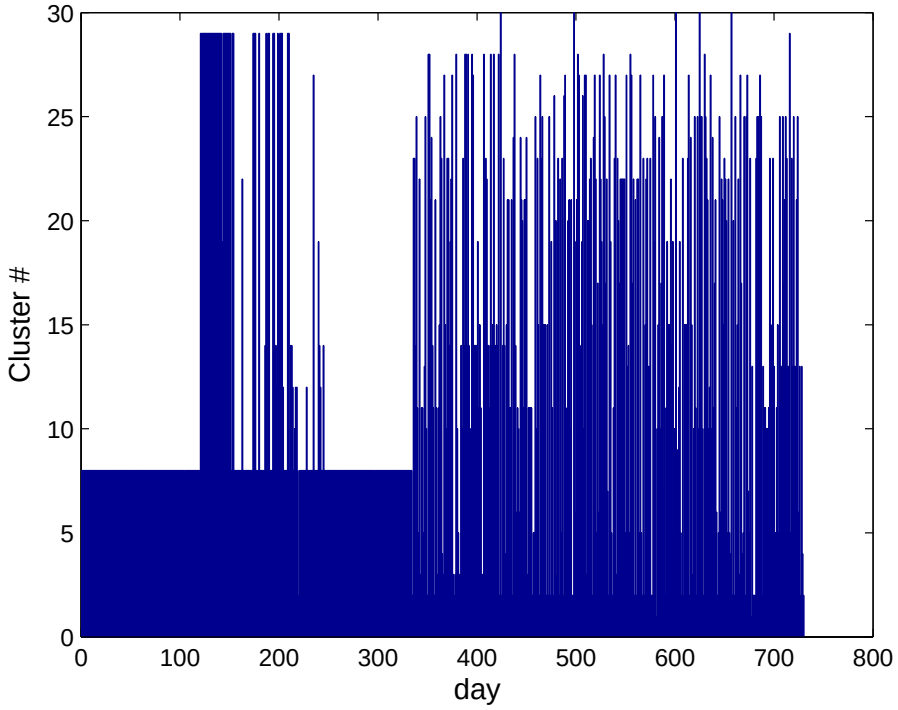
No.11 in 1st Monday, 2010), then we can easily make a judgement that some load shape change happens at this day.

So the basic idea of detecting load shape change in this method is trying to transform load shape change into classification or clustering result change based on its daily AMR data segment. Therefore, we can focus on finding the moving of these classification or clustering results.

In practice, we can firstly implement the K-means classification algorithm based on the temperature normalized annual AMR data of one specific year (e.g. year 2009), after it K centroids can be obtained to represent K different typical annual load profiles. Then based on these K annual load profiles, we can divide every annual load profile into 365 daily segments to get $K\times365$ daily load profiles. Next if we want to observe the customer daily behavior change in one specific year, it can be done by dividing the customer's temperature normalized AMR measurements in this year into the same amount (i.e. 365) of segments. Every temperature normalized daily AMR segment in this year will be compared to all the K typical daily load profiles to be decided which cluster it should belong to. The Euclidean distance is used as comparing metric and the clustering results of daily load profiles are shown in Fig.6.5. Now we can observe the intra-year daily behavior variability of one specific customer by observing everyday clustering results during one complete year. And we can also observe the daily behavior variability between two years (i.e. 365×2 days in Fig.6.5b and $365\times24\times2$ hours Fig.6.5a) by comparing clustering results at the similar date.



(a) Temperature normalized AMR measurements for customer No.1496 in two years



(b) Daily load clustering results for customer No.1496 in two years

Figure 6.5: Daily load change detection based on clustering results

Unfortunately most kinds of load shape changes are not significant enough to affect the clustering results and just can make small change inside one cluster. Thus this kind of small change can not be detected by the basic K-means algorithm and some improvement should be developed to make this load profile clustering method more accurate. In following, it is found that Fuzzy C-means can work well for this issue and give better load shape change detection results. This method will be discussed in detail in Chapter 7.

7. WEEKLY LOAD PROFILING WITH FUZZY C-MEANS

The aim of this chapter is trying to introduce a method which can detect those behavior changes that are not big enough to affect reclassification results. It means that when some customers change their electricity consumption behavior but still remain within the same clustering group, it is not possible to use the clustering index mentioned above to indicate the temporary change or small random change. To some extent, the idea of this method is trying to decompose the customer behavior into several bases, where we choose clustering centroids as bases. After the bases are given, we assign a coefficient to every base (i.e. clustering centroid) to measure the grade of how much one specific customer behavior matches against this base. This idea can be interpreted as follow:

$$\text{Customer Behavior} = a_1 \times \text{cluster}_1 + a_2 \times \text{cluster}_2 + a_3 \times \text{cluster}_3 + \ldots + a_K \times \text{cluster}_K$$

where:

a_i: is the coefficient to measure how much the customer behavior matches cluster i,
cluster_i: is the ith typical customer behavior base.

The subscript K depends on the number of clusters which we choose and K is chosen to be 30 in following. For different years, we can build such different representations of this specific customer behavior by assigning different sets of a_i values. Actually these a_i values are decided by membership in Fuzzy C-means algorithm, which is introduced in Chapter 3. Then by comparing these different sets of a_i values in different years, we can detect the behavior change through observing the change of a_i values during one specific time interval. The flowchart of the whole method is shown in Fig.7.1.

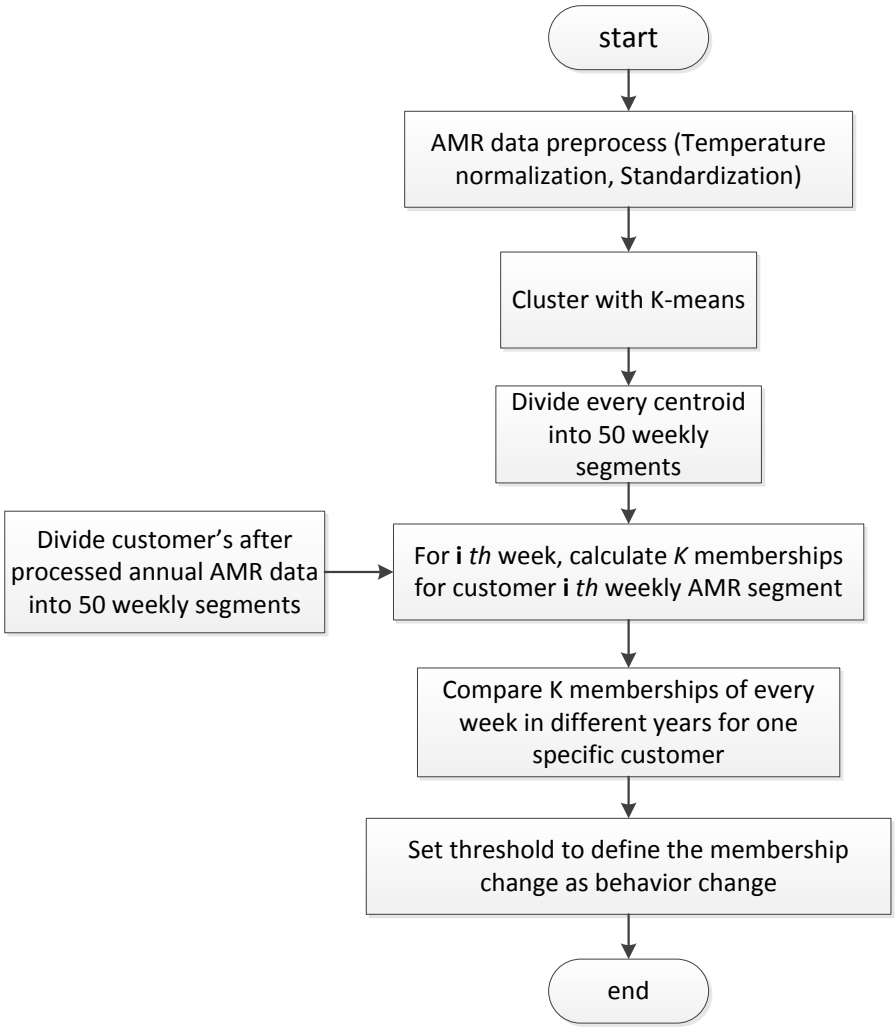


Figure 7.1: Flowchart of load shape change detection method

7.1 Weekly load profiles and membership

As shown in Fig.7.1, this method mainly involves getting the weekly load profile (i.e. centroid segment) and customer weekly behavior (AMR weekly segment) to calculate K membership for every week of different years. In order to obtain these weekly load profiles and membership, this method can be divided into three steps.

7.1.1 Clustering annual AMR measurements to obtain centroids

The first step is clustering all the non-empty customers' hourly measured AMR data in one complete year to obtain K centroids using conventional K-means algorithm.

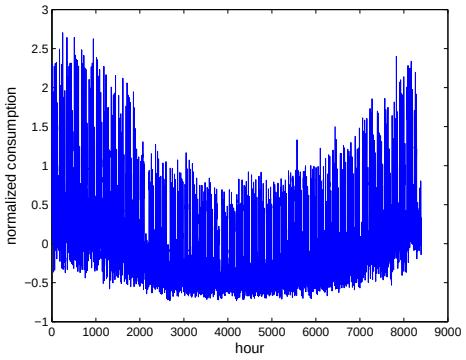
Those examples here come from KSAT case but this methodology can work for any data set. Before running clustering, the AMR data of every customer should be preprocessed by temperature normalization and normalized to make sure the obtained centroids are mostly at the same level (i.e. standard normal distribution) but with different shape. Function *zscore* in MATLAB is used to achieve the following normalization:

$$\frac{\mathbf{x}_i - mean}{standard \quad deviation} \tag{7.1}$$

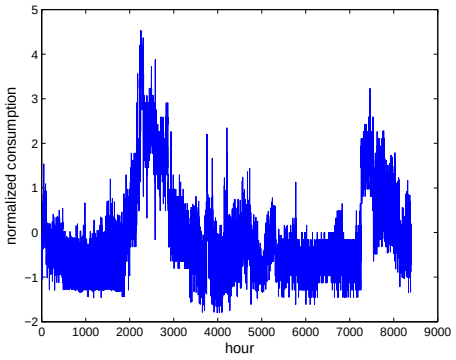
where:

$\mathbf{x}_i$ is: ith customer's AMR data vector.

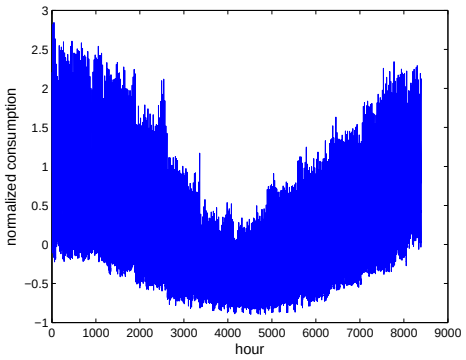
Here every centroid can be called "annual load profile" because its dimension is 8760 (i.e. 365 days×24 hours). After it, we will have a group of 30 (K=30 is chosen to be the number of clusters here) "annual load profiles", some of them are shown in Fig.7.2.



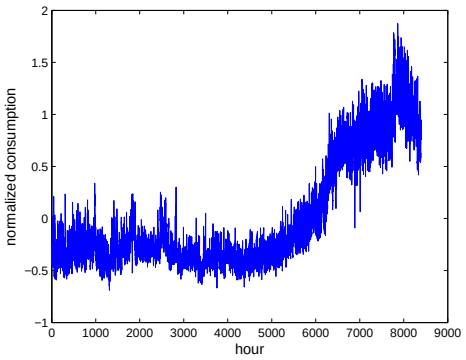
(a) The 1st annual load profile



(b) The 2nd annual load profile



(c) The 5th annual load profile



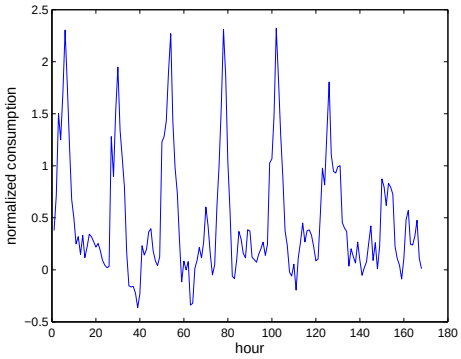
(d) The 30th annual load profile

Figure 7.2: Annual load profiles (i.e. centroids) obtained from running K-means

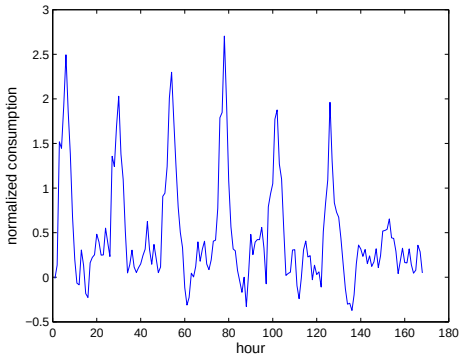
7.1.2 Dividing annual load profiles into weekly load profile

The second step is the so-called "weekly load profiling", where each of these 30 annual load profiles is divided into 50 weekly segments. Every segment is 168 dimensional (i.e. 7 days×24 hours). Thus 50 integral weeks can always be obtained from one complete year. It should be pointed out that when we divide every annual load profile into weekly segments, every weekly segment should begin from Monday to Sunday. That means we cannot simply divide the whole year into 365/7 weeks.

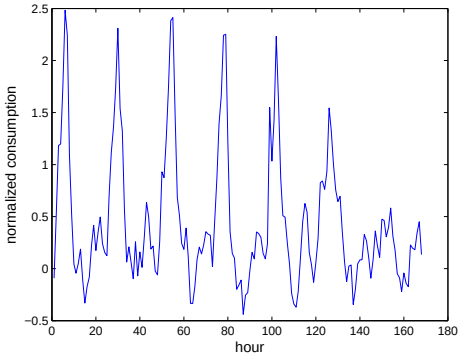
The Fig.7.3 shows weekly load profiles extracted from 1st annual load profile. To emphasize this process, it can be interpreted in this way that we divide every annual load profile into 50 weekly load profiles and actually every weekly load profile is just part of the annual load profile. Hence in total, we will have 1500 weekly load profiles (i.e. 30 centroids × 50 weeks). Every one of K=30 annual load profiles (i.e. centroids) has its own corresponding 50 weekly load profiles (i.e. weekly centroid segment).



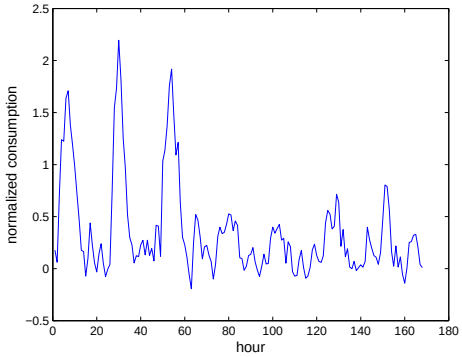
(a) The 1st weekly load profile



(b) The 2nd weekly load profile



(c) The 5th weekly load profile



(d) The 50th weekly load profile

Figure 7.3: The weekly load profiles extracted from 1st annual load profile

7.1.3 Obtaining membership based on weekly load profiles

Now we also need to divide customer annual AMR measurements in different years into 50 weekly AMR data segments for every year (e.g. year 2009 and year 2010). Before this, customer annual AMR measurements also go through the data preprocess of temperature normalization and standardization. It is worth mentioning here that the customer's 1st weekly AMR data segment always begin from the first Monday of that year. Because for instance it is unreasonable to directly compare 1st day of year 2009 (Thursday) to 1st day of year 2010 (Friday). Customer may have totally different behavior in different day types. For every customer, its electricity consumption behavior of the whole year is divided into 50 weekly electricity consumption behavior. In other words, the annual electricity consumption behavior of one specific customer is represented by these 8400 (i.e. 50 weeks $\times$ 168 hours/week) dimensional data. The rest 360 hours of data is abandoned. The after preprocessed annual AMR data for customer No.444 in different years (i.e. 2009 and 2010) are shown in Fig.7.4 and Fig.7.5.

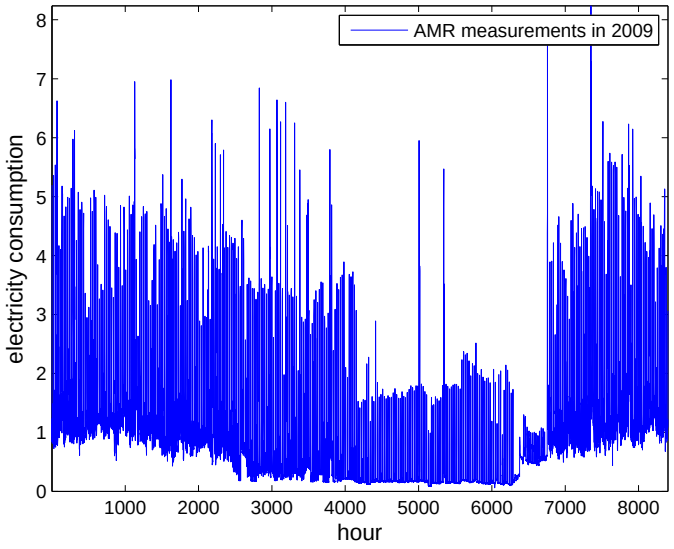


Figure 7.4: The annual 8400 dimensional AMR data for customer No.444 in 2009

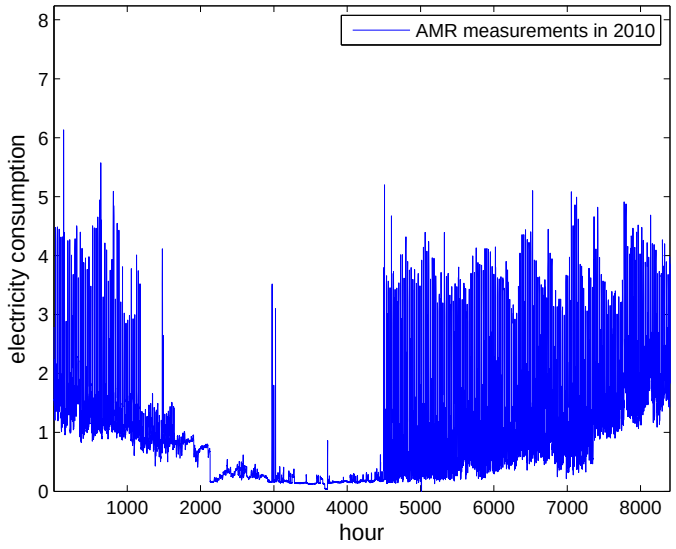


Figure 7.5: The annual 8400 dimensional AMR data for customer No.444 in 2010

Now we can use the formula (3.3) in Chapter 3 from the Fuzzy C-Means algorithm (FCM) to obtain membership for every weekly AMR data segment of one specific customer in different years. It is done by comparing every weekly AMR data segment in different years to all the K weekly load profiles in every weekly time interval. For each week, the customer weekly consumption behavior (i.e. weekly AMR data segment) will be represented by K memberships as shown by 30 different colors in Fig.7.6.

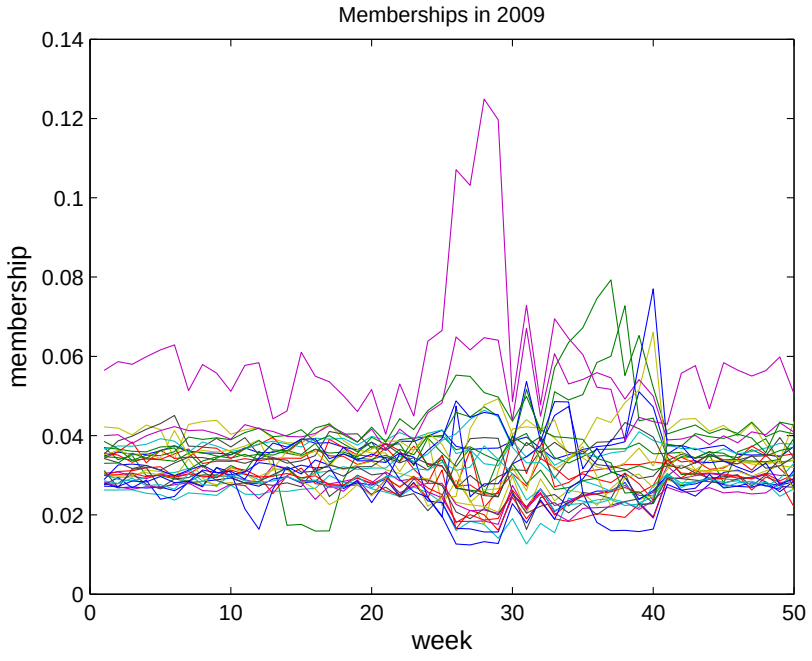
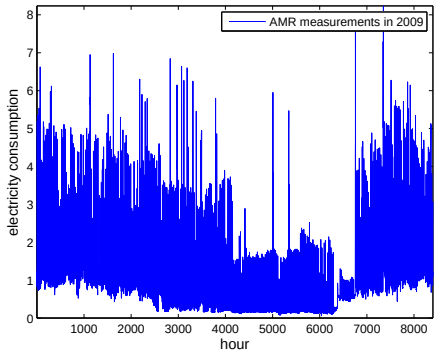


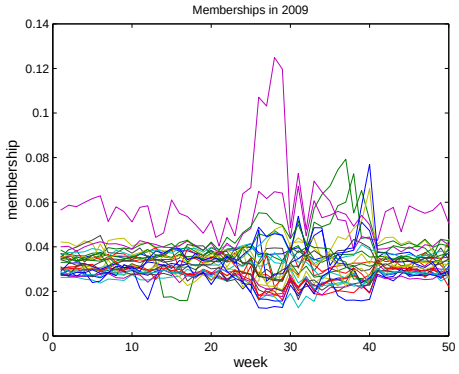
Figure 7.6: Membership for customer No.444 in year 2009

Here we do not use the complete the Fuzzy C-Means algorithm (FCM) to do clustering. The aim is just to get membership information in FCM. We also tried using the complete Fuzzy C-Means algorithm instead of the K-means algorithm to do clustering at step one to achieve annual load profiles. But we found in fact it is not necessary to do so since only the centroids information is needed at step one. Actually, these memberships are calculated to determine how much degree this customer's weekly AMR data segment (i.e. weekly consumption behavior) matches the corresponding K weekly load profiles in every week. It also should be pointed out that these membership values are quite low maybe due to the reason that it is hard for such high dimensional data example to completely belong to any one of those K clusters.

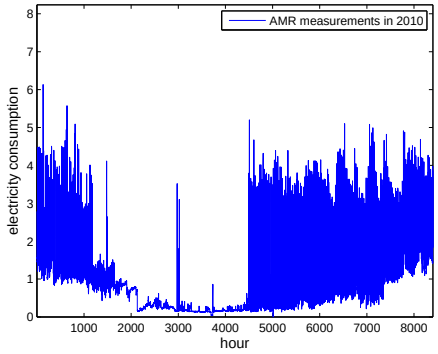
In summary, every one of these 50 weekly AMR data segments for one specific customer in different years will be matched against everyone of K weekly load profiles in every weekly time interval to see how much this customer's weekly consumption behavior looks like every weekly load profile. These weekly load profiles are actually extracted from annual load profiles (i.e. centroids). The degree of such matching is exactly measured by the membership calculation method from Fuzzy C-means algorithm and K is chosen to be 30 here. During the whole year, this kind of comparison or matching will go through total 50 times to make sure every weekly consumption behavior of one specific customer has 30 different memberships. 50 groups of 30 memberships are obtained in one complete year. For different years, we will get different sets of memberships to make a comparison. Based on this, some conclusion or judgement about whether one customer's behavior changes or not in different years will be obtained. It can be obviously observed in following Fig.7.7.



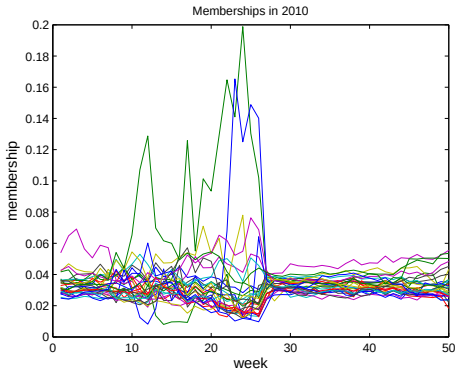
(a) Consumption behavior for customer No.444 in 2009



(b) Memberships for customer No.444 in 2009



(c) Consumption behavior for customer No.444 in 2010



(d) Memberships for customer No.444 in 2010

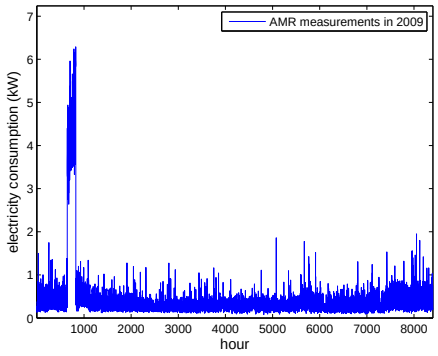
Figure 7.7: Memberships for customer No.444 in 2010

From the above figures, some obvious AMR data changes and membership changes can be observed in these two different years.

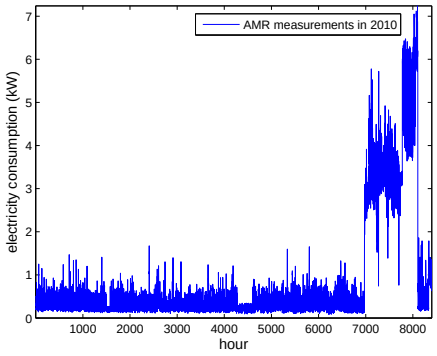
7.2 Change detection based on memberships

When we have a bunch of membership curves, the intuitive way to think of detecting the change based on these memberships is to calculate the absolute difference of every cluster membership in every specific week. Then a threshold is needed to be set to determine how much degree the membership change can be seen as the customer behavior change. That helps us judging a customer behavior change by observing whether or not the amount of membership difference in some weeks exceeds the threshold. Here we set the threshold as constant 0.1 and sum up the 20 most significant absolute difference of membership in every week. It is realized by function *sort* in MATLAB and shown in Fig.7.8e.

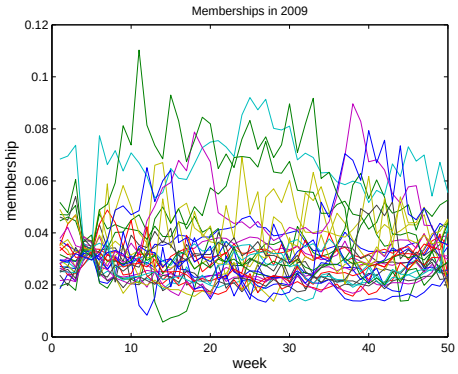
In Fig.7.8, it is easy to observe the huge difference in Fig.7.8e mainly appear around week 5 and week 45. According to the calculation $4 \times 7 \times 24 + 1 = 673$ hours and $44 \times 7 \times 24 + 1 = 7393$ hours, it can be easily checked from Fig.7.8a



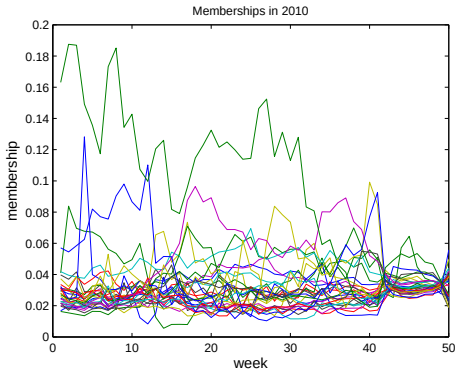
(a) AMR measurements for customer No.1836 in 2009



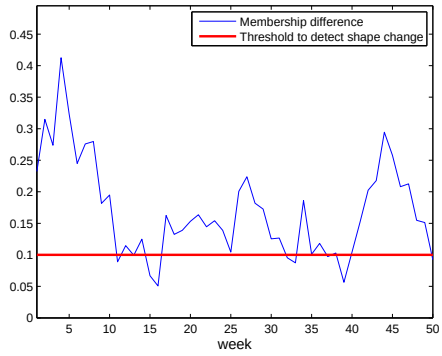
(b) AMR measurements for customer No.1836 in 2010



(c) Membership for customer No.1836 in 2009



(d) Membership for customer No.1836 in 2010



(e) Membership change between two years

Figure 7.8: Change detection based on memberships of two different years

and Fig.7.8b that some obvious behavior difference indeed appear around hour 673 of year 2009 and hour 7393 of year 2010. Thus some customer behavior change can be said to happen around 5th week in the year 2010 compared with the year 2009. And it should also be noticed here that as mentioned before, this kind of hour counting begins from the first Monday of every specific year. So some trivial adding of previous abandoned hours is needed to find the correct time of this change

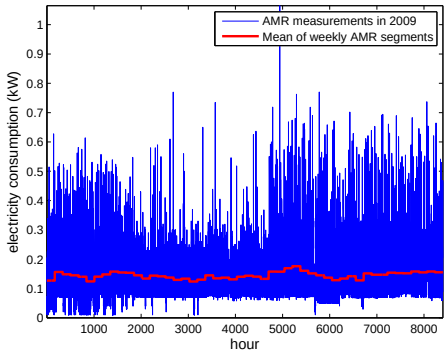
happening in one specific year. On the other hand, due to the limit that this method compresses the hourly information into weekly information, actually just the approximate time of change happening can be obtained. And the crossing of the membership difference to the threshold should also continue several weeks to make sure it is not just caused by some temporary abnormal behavior.

7.3 Load shape change detection example

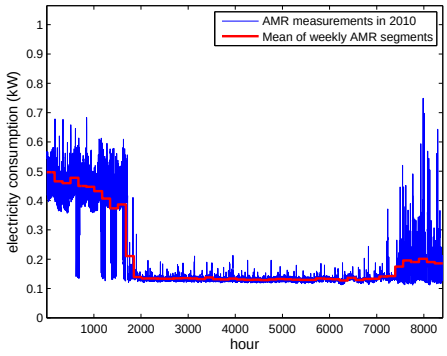
This weekly load profiling method is mainly used to detect customer's load shape change, especially those changes that can not be detected by the load level change detection method introduced in Chapter 5. For instance, in Fig.7.9 the customer No.2455 obviously has completely different load shapes in the year 2009 and the year 2010, but since its consumption level in these two years are almost the same, this kind of change can just be detected by weekly load profiling method.

Through comparing Fig.7.9e with Fig.7.9f, we can easily find that load level change method does not inform any difference in the middle part of these two years due to this customer No.2455 has almost the same electricity consumption in the middle part of two different years. But the obvious load shape change still cause the change of membership in these two years. Thus this kind of shape change will be detected by the weekly profiling method.

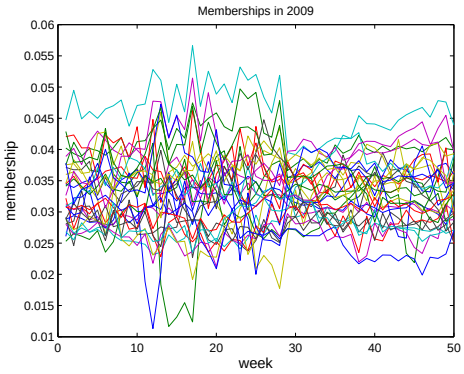
In summary, based on these membership curves we transform the change detection of high dimensional AMR measurements to the change detection of these low dimensional membership curves. The benefit of such transform is not only about dimension reduction but also helpful to reduce the effect of random variance noise from some measurement outlier or temporary abnormal behavior. All of them are extremely harmful for analyzing the customer behavior change.



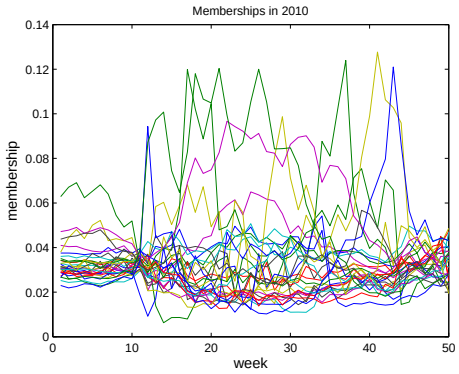
(a) AMR measurements for customer No.2455 in 2009



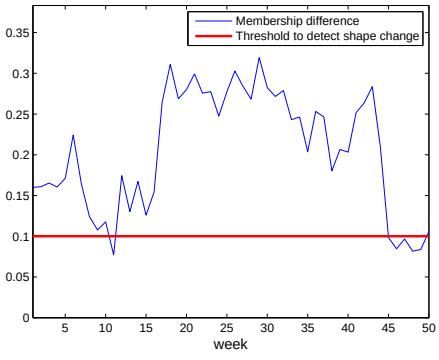
(b) AMR measurements for customer No.2455 in 2010



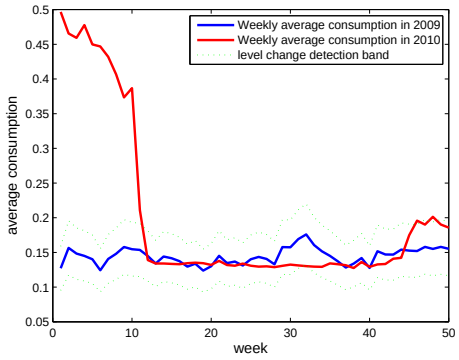
(c) Membership for customer No.2455 in 2009



(d) Membership for customer No.2455 in 2010



(e) Load shape change based on membership between two years



(f) Load level change between two years

Figure 7.9: Weekly load profiling to detect load shape change for customer No.2455

8. TEST CASES AND METHOD VALIDATION

This chapter aims to test the weekly load profiling method with some typical behavior change cases. In general, based on observing hundreds of electric customers' behaviors, it is fair to say almost every customer changes electricity consumption routine more or less between different years. This phenomenon is reasonable considering that it is impossible for human behavior to repeat without any habit change. Even for industrial customers, socio-economic conditions are so different in several years and often affect their electricity consumptions. The aim of this thesis is to detect those obvious big enough or abnormal changes rather than focusing too much on regular variations. So some typical kinds of electric behavior changes, such as the change of heating solution and the change of periodicity, are tested.

8.1 Artificial test data

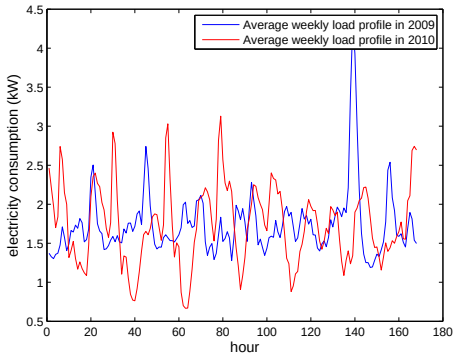
In order to test the validity of the method proposed in this thesis, some virtual customers can be created through combining different customers' behavior in different years. Firstly, we find 928 customers with no behavior change at all, each of them has almost the same load level and load shape in two different years. Then we combine the consumption behavior of customer No.1-No.800 in 2009 with the consumption behavior of customer No.101-No.900 in 2010 to create customers with load shape change. In order to make sure there is shape change for every one of these 800 artificial customers, those consumption behaviors coming from the same cluster by coincidence are removed. Then 709 artificial customers with load shape change are left for test purpose. We also multiply every hourly consumption of customer No.1-No.800 in 2010 by an absolute value of a factor, which obeys standard normal distribution, to create customers with load level change. In this way, these 709 load shape change customers and 800 load level change customers are tested as follows:

Table 8.1: Artificial customers with behavior changes detected

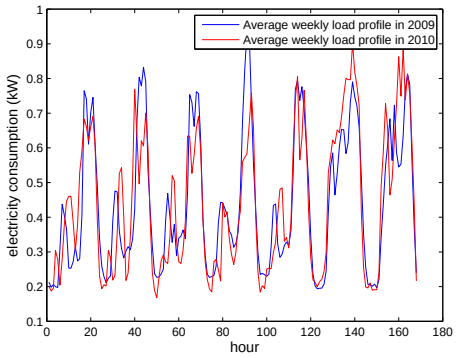
Customer type	#Cus.	#Cus. with change detected	Per.
Shape change Cus.	709	649	91.5%
Level change Cus.	800	704	88.0%

The artificial customer examples in Fig.8.1 are average weekly load profiles of some different type customers with load level change or load shape change. Some of

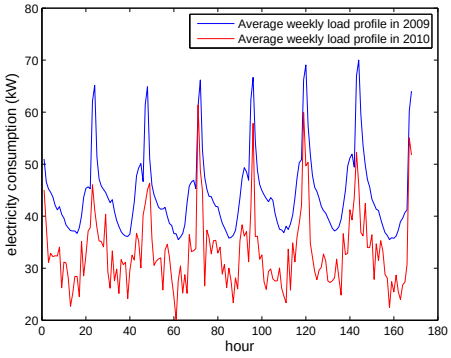
these changes are successfully detected and some of them are not.



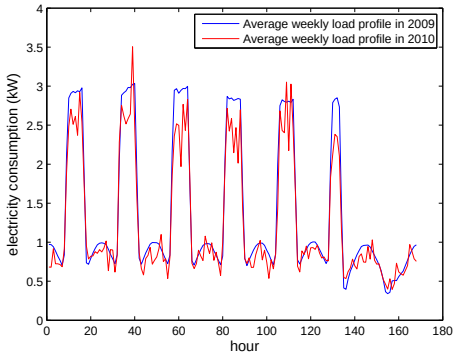
(a) Artificial customer with load shape change detected



(b) Artificial customer with load shape change but not detected



(c) Artificial customer with load level change detected



(d) Artificial customer with load level change but not detected

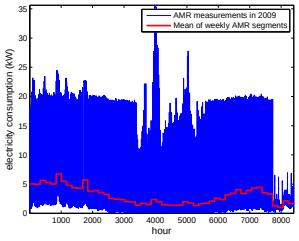
Figure 8.1: Average weekly load profiles of different type artificial customers

It is fair to say most of the artificial customer behavior changes are successfully detected by the method proposed in this thesis. But it does not work so well for those customers who often have many consumption peaks during any time interval. Those peaks or sparks bring so much uncertainty to either the load shape or load level, involving a lot of outlier when we measure the distance in Euclidean space. Fortunately most customers have just limited amount of peaks, so this method can achieve about 90% accuracy in this detection task based on the artificial data set. In next, we will implement this method into real data sets to obtain some meaningful results.

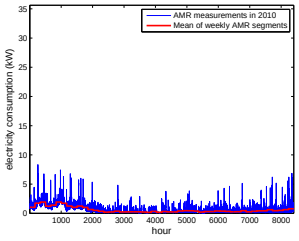
8.2 KSAT data

The following test case comes from KSAT data set introduced in Chapter 4. Both the load level change detection method and the load shape change detection method are implemented here.

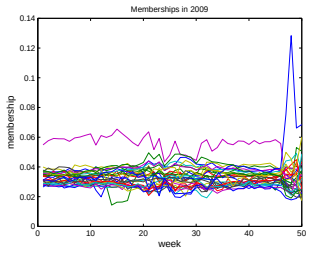
Fig.8.2 shows a customer No.1496 with class number 4. According to the table A.1 in the appendix, it belongs to the type of "Housing+storage electric heating". It is easy to find that this customer has completely different load levels in two different years. The change begins almost from the beginning of the year 2010 compared with the year 2009.



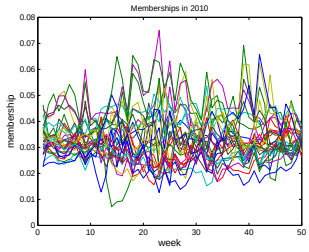
(a) AMR measurements in 2009



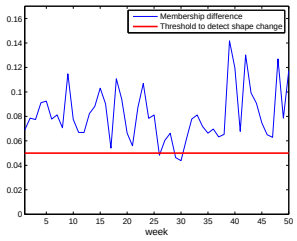
(b) AMR measurements in 2010



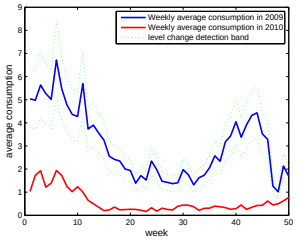
(c) Membership in 2009



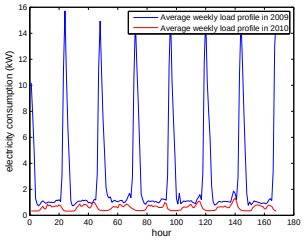
(d) Membership in 2010



(e) Load shape change based on membership between two years



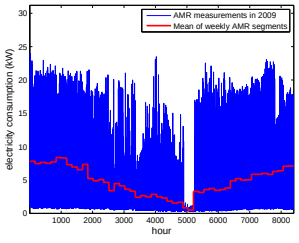
(f) Load level change between two years



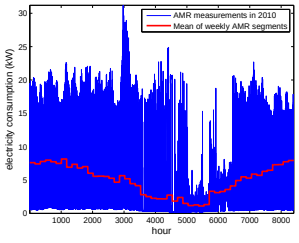
(g) Average weekly load profiles for customer No.1496

Figure 8.2: Change detection for customer No.1496 in year 2009 and 2010

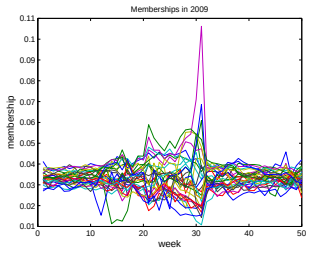
Another customer No.2771 is labelled with class number 1, so it has typical "Housing" load curve shape. The detection result is shown in Fig.8.3.



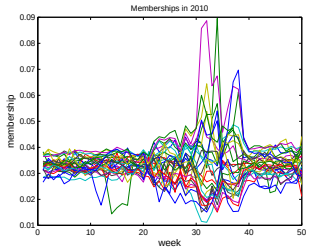
(a) AMR measurements in 2009



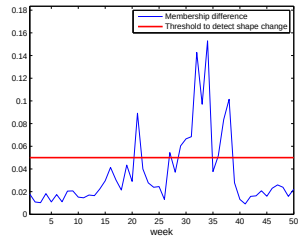
(b) AMR measurements in 2010



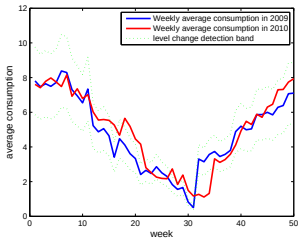
(c) Membership in 2009



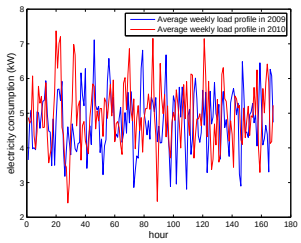
(d) Membership in 2010



(e) Membership difference between two years



(f) Load level change between two years



(g) Average weekly load profiles for customer No.2771

Figure 8.3: Change detection for customer No.2771 in year 2009 and 2010

By observing the cross point region in Fig.8.3e and Fig.8.3f, the load shape change is mainly supposed to happen during 33th week to 38th week, and the load level change also happens around 18th week to 22th week. It can also be further checked by directly comparing the AMR measurements of this customer in the year 2009 and the year 2010. Because the load level change does not appear more than 10 weeks in one year period, perhaps it is fair to conclude this customer just has temporary change.

In order to find how many customers in KSAT data set change their behavior from the year 2009 to the year 2010, we set some criteria for load level change detection and load shape change detection respectively. For load level change detection, if a customer load level at one year exceeds the $\pm 125\%$ weekly average consumption band of another year at least 10 times, we will conclude that this customer behavior changes in different years. For load shape change detection, if the sum of membership differences of one customer exceeds the threshold 0.01 at least 10 times, then we can

conclude that this customer behavior change happens. In other words, if there are at least 10 weeks when the weekly loads are different in two years, the first change happening week will be recorded.

Firstly, we run the load level detection method introduced in Chapter 5 over all the 3577 customers in KSAT data set. The result is shown in table 8.1 and the weekly information when their load level changes happen for the first time is shown in Fig.8.4. This histogram shows that if one customer has load level change, at which week these level changes begin to appear. In most cases, if one customer changes his consumption behavior from the whole year perspective, it is very likely that some change will appear even in the first week.

Table 8.2: Number of customers who have load level change

	#customers	percentage
load level change	2323	64.9%
no load level change	1254	35.1%

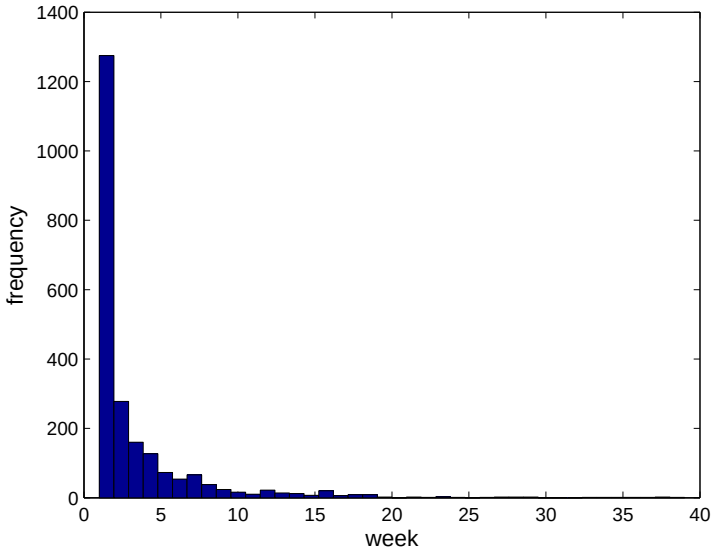


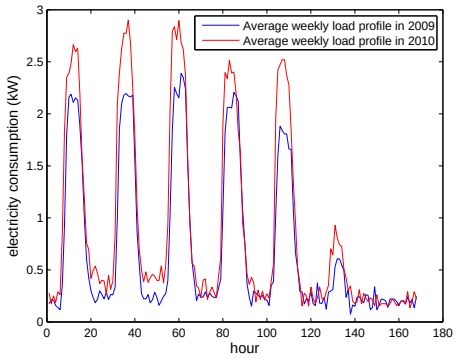
Figure 8.4: Histogram for week information of load level changes happening for the first time

Secondly, we run the weekly load profiling method over all the 3577 customers in KSAT data set and compare the result with the reclassification method. Results are listed in the following table.

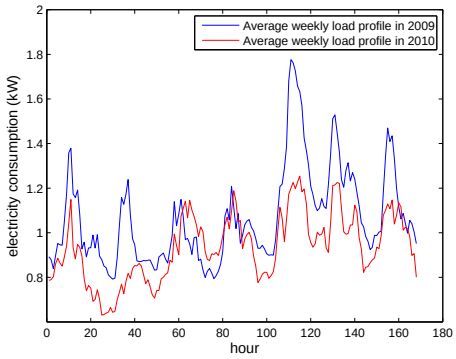
Table 8.3: Number of customers who have load shape change

Change detected	#customers	percentage
only by reclassification	346	9.7%
only by weekly load profilling	1146	32.0%
by both mehods	603	16.9%
no change	1482	41.4%

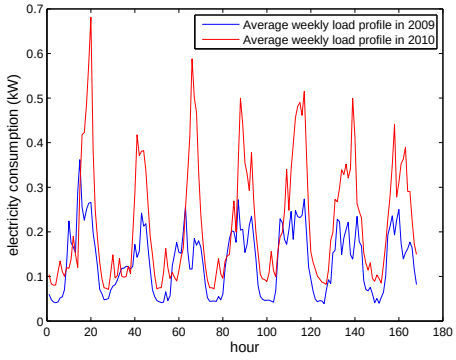
Some typical behavior change customers are also shown in Fig.8.5.



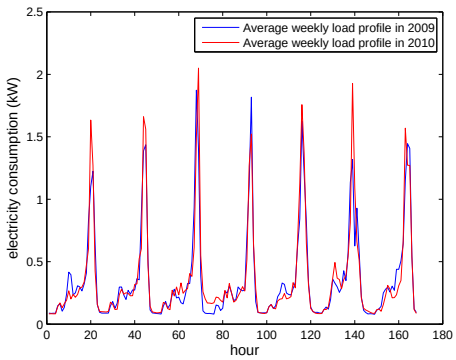
(a) Average weekly load profile of customer with change only detected by reclassification method



(b) Average weekly load profile of customer with change only detected by weekly load profilling method



(c) Average weekly load profile of customer with change detected by both methods



(d) Average weekly load profile of customer without change

Figure 8.5: Different type customers in KSAT data set with change detected

If some customer behavior changes are detected by the weekly load profiling method, the time information regarding at which week the change happens for the first time can also be offered as shown in Fig.8.6.

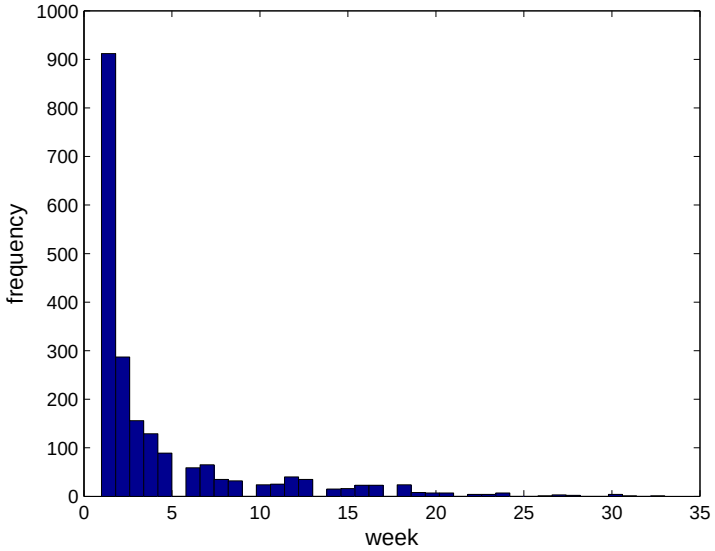


Figure 8.6: Histogram for week information of load shape changes happening for the first time

8.3 Elenia data

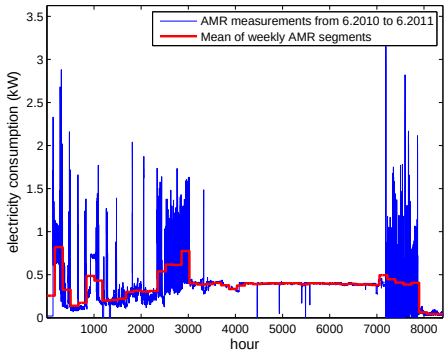
Elenia data set has 7398 non-empty customers in total. The following Fig.8.7 shows the behavior of customer No.1970 in June 2010 to June 2012 from Elenia data set. We can observe that load level change detection matches with the load shape change detection as shown in Fig.8.7f and Fig.8.7e.

The result of checking all the 7398 customers with the load level change detection method and the load shape change detection method is shown in the following table.

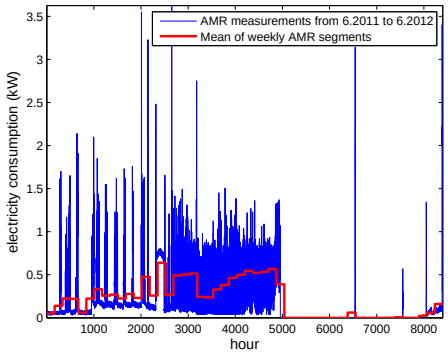
Table 8.4: Number of customers with behavior change

Change type	#customers	percentage
only load level change	1402	19.0%
only load shape change	265	3.6%
both changes	3969	53.6%
no change	1762	23.8%

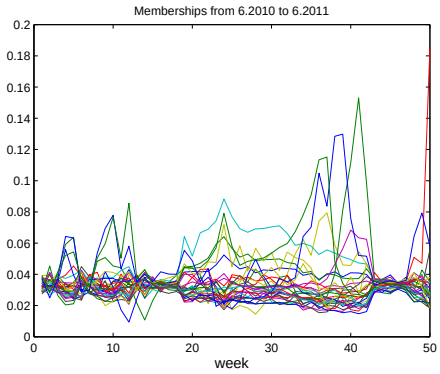
From this table, we can see most customers have both load level change and load shape change. In other words, it is hard to separate load level change from load shape change, either of them in most cases will introduce the other. That is why there are 1315 customers only having load level change but just 301 customers only having load shape change. It means most load shape changes will also bring some load level changes during some certain time intervals. And in total about 75% customers change their behaviors more or less during different years.



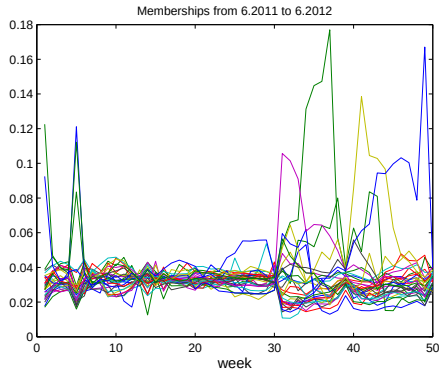
(a) AMR measurements of customer No.1970 from June 2010 to June 2011



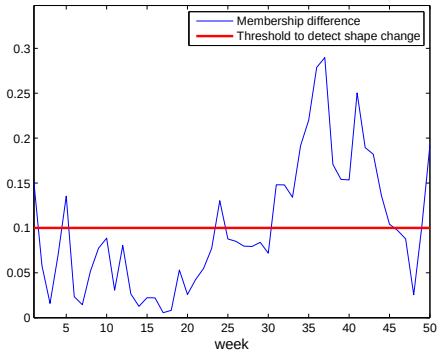
(b) AMR measurements of customer No.1970 from June 2011 to June 2012



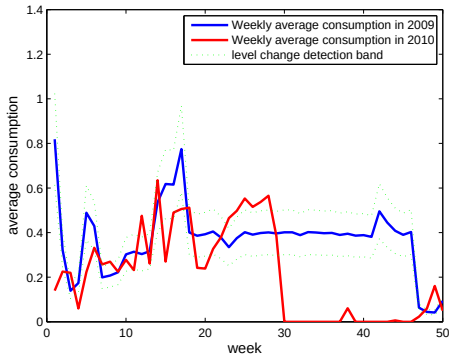
(c) Membership of customer No.1970 from June 2010 to June 2011



(d) Membership of customer No.1970 from June 2011 to June 2012



(e) Membership difference between two years



(f) Load level change between two years

Figure 8.7: Change detection for customer No.1970 from June 2011 to June 2012

9. CONCLUSION

The customer behavior change detection is not an easy problem since most customer behaviors are quite irregular and accompanied with a number of random variations. Especially when we want to know the accurate time information regarding these changes happening, it becomes even harder. Here in this thesis, the main goal is trying to find a method to separate the customer load level change and load shape change and detect them respectively. So the weekly load profiling method is proposed based on the work of customer classification to detect the load shape change. For load level detection, a fixed weekly time window is formed to detect the consumption level difference. Both of them depend heavily on our own definition of customer change, namely the customer behavior has some difference in different years. Actually, it should be called between-year change, one of many various definitions of behavior changes, such as intra-year change and abrupt change. On the other hand, since the customer behavior is not stable as we emphasized, the weekly profiling method developed in this thesis can just offer the weekly time information regarding at which week the behavior change happens compared with the previous year. It is hard to offer any further daily or hourly information due to the limit of classification methods.

In general, this method works well for large customers with obvious change but may neglect some peak change of small customers. It can also be used in the situation that reclassification of the same customer in different years produces no changed result and during some certain weekly time intervals, some obvious changes still happen. However, this method can not be used to detect some small and temporary changes that last just a few hours or days. Some other methods which take these AMR data as time series with intrinsic patterns instead of high dimensional vectors should be developed in future to detect more trivial electricity customer behavior change. Especially, when we want to measure the similarity of two times series, Euclidean distance is a good but may not be the perfect way to do such measuring if our focus is on the time series intrinsic patterns. Some technique such as dynamic time warping can help solving the time shifting and salient points shifting problems, which are difficult for using Euclidean distance. Other methods like fuzzy logic, artificial neural network, chaos theory and soft computing or computational intelligence may be possible ways to solve this problem at root.

REFERENCES

[Adams 2007] Adams, R. P. and MacKay, D. J., Bayesian online changepoint detection. Technical report, University of Cambridge, Cambridge, UK, 2007

[Brockwell & Davis 2009] Brockwell, P. J. and Davis, R. A. (2009). Time Series: Theory and Methods (2nd ed.). New York: Springer. p. 273.

[Bishop et al. 2006] Bishop, Christopher M., Pattern Recognition and Machine Learning, U.S. 2006, Springer press, 424 p.

[Chicco et al. 2006] Chicco, G. , Napoli, R. , and Piglione, F. , Comparison among clustering techniques for electricity customer classification, IEEE Trans. Power Syst., vol.21, no.2, pp.933-940, May 2006.

[Chen et al. 1996] Chen, C.S, Hwang, J.C, Huang, C.W, Tzeng, Y.M and Chu, M.Y., Application of Load Survey System at Taipower, IEEE Transaction on Power Delivery, vol.11, 1996

[Duda et al. 2000] Duda, Richard O., Hart, Peter E., Stork David G., Pattern Classification, 2nd Edition, UK 2000, Wiley press, 526 p.

[Darby, 2010] Darby, S., Smart metering: What potential for householder engagement ?, in Building Research & Information. Evanston, IL: Routledge, 2010, vol.38, pp. 442-457.

[Degefa 2010] Degefa, Merkebu Zenebe. 2010. Energy Efficiency Analysis of Residential Electric End-Uses: Based on Statistical Survey and Hourly Metered Data. Master of Science Thesis. Aalto University. 70 p.

[Espinoza et al. 2005] Espinoza, M., Joye, C., Belmans, R., De Moor, B., Short-Term Load Forecasting, Profile Identification, and Customer Segmentation: A Methodology Based on Periodic Time Series, IEEE Transaction on Power System, vol.20, issue.3, pp.1622-1630, Aug. 2005.

[Filho et al. 1991] Filho, Do Coutto, Leite, Da Silva, Arienti, V. L., Ribeiro, S. M. P . Probabilistic Load modeling for Power System Expansion Planning. IEE, Third International Conference of Probabilistic Methods Applied to Electric Power System, 1991. P. 203-207.

[Figueiredo et al. 2005] Figueiredo, Vera, Rodrigues Fatima, Vale Zita A., Borges Gouveia. An Electric Energy Characterization Framework based on Data Mining Techniques, IEEE Transactions on Power Systems, vol. 20, nr. 2, pp. 596-602, May 2005.

[Herman et al. 1993] Herman, R., Kritzinger J. J . The statistical description of grouped domestic electrical load currents, Electric Power System Res. 1993. Vol. 27. P. 43-48.

[Hahn et al. 2009] Hahn, H., M.Nieberg, S., Pickl, S., Electric load forecasting methods: Tools for decision making, European Journal of Operational Research, vol.199, issue3, pp.902-907, Dec. 2009.

[Jain et al. 1999] Jain, A. K., Murty, M. N., Flynn, P. J., Data clustering: a review, ACM Computing Surveys (CSUR). Vol. 31 Issue 3, pp.264-323, Sept. 1999.

[Koreneff 2009] Korenfeff, G., Ruska, M., Kiviluoma, J., Shemeikka, J., Lemstrom, B., Tiina K.A., Future development trends in electricity demand, VTT Technical Research Centre of Finland, Espoo, 2009, pp.8-9.

[Laitinen et al. 2011] Laitinen, A., Ruska,M. and Koreneff, G., Impacts of large penetration of heat pumps on the electricity use, SGEM project report, 2011, Espoo, pp.15

[Meldorf et al. 2007] Meldorf, M., Taht, T. and Kilter, J. Stochasticity of the electrical network load, Oil Shale;2007 Supplement, Vol. 24, p225, April 2007.

[Mutanen et al. 2008] Mutanen, A., Repo, S. and Järventausta, P., AMR in Distribution Network State Estimation, presented at the 8th Nordic Electricity Distribution and Asset Management Conf., 2008, Bergen, Norway.

[Mutanen 2010] Mutanen, A., Customer classification and load profiling based on AMR measurements, research report, Tampere University of Technology, 2010, pp.3-4.

[Mutanen 2011] Mutanen, A., Using AMR measurements in load profiling and network calculation, Load and response modeling workshop in project SGEM, 10 November 2011, Kuopio.

[Mutanen et al. 2011] Mutanen, A., Ruska, M., Repo, S., and Järventausta, P., Customer classification and load profiling method for distribution systems, IEEE Trans. Power Del., vol. 26, no. 3, pp. 1755-1763, Jul. 2011.

[Mutanen 2013] Mutanen, A., Methods for using AMR measurements in load profiling, SGEM project report, 2013, pp.11-12.

[Neimane et al. 2001] Neimane, V. Distribution Network Planning Based on Statistical Load Modeling Applying Genetic Algorithms and Monte-Carlo Simulations. IEEE, Porto PowerTech Conference, 2001. Vol. 3. 5 pp.

[Smart Grid 2012] Webpage. [accessed on 21.7.2014]. Available at: http://en.wikipedia.org/wiki/File:Smart_Grid_Function_Diagram.png

[Seppälä 1996] Seppälä, A., Load research and load estimation in electricity distribution, Ph.D. dissertation, Helsinki Unv. Technol, Espoo, Finland, 1996.

[Stephen et al. 2014] Stephen, B., Isleifsson, F. R., Galloway, S., Burt, G.M. and Bindner, H.W. Online AMR Domestic Load Profile Characteristic Change Monitor to Support Ancillary Demand Services, Smart Grid, IEEE Transactions on, vol. 5, pp. 888 - 895 , March 2014.

[Stephen&Mutanen et al. 2014] Stephen, B., Mutanen A., Galloway, S., Burt, G. and Jarventausta, P., Enhanced Load Profiling for Residential Network Customers, IEEE Trans. Power Delivery, vol. 29, pp. 88-96, Feb 2014.

[Timothy 2011] Timothy, C. U., Statistics in Plain English (3rd. ed), Routledge press, 2011, pp. 105

[U.S. Dep.Energy 2012] U.S. Department of Energy, Smart Grid / Department of Energy. Retrieved 2014-06-18. http://energy.gov/oe/services/technology-development/smart-grid

[Wang 2009] Wang, J. J., An ARMA Cooperate with Artificial Neural Network Approach in Short-Term Load Forecassting, Natural Computation, 2009.ICNC 09. Fifth Internatinal Conference on, Tianjin, 2009

[Yao and Steemers 2005] Yao, R. and Steemers, K., A Method of Formulating Energy Load Profile for Domestic Buildings in the UK. Energy and Buildings. New York:Elsevier, 2005, vol.37, pp.663-671.

A. APPENDIX

Table A.1: Customer types in KSAT data set

Class	Power factor	Type of customer
1	0.96	Housing
2	0.96	Housing + direct electric heating
3	0.96	Housing + partial storage electric heating
4	0.96	Housing + storage electric heating
5	0.96	Detached house, no electric heating, electric sauna stove
6	0.96	Detached house, no electric heating, no sauna stove
7	0.96	row house/apartment, no electric heating, electric sauna stove
8	0.96	row house/apartment, no electric heating, no sauna stove
9	0.96	Agriculture + housing
10	0.96	Agriculture + housing + electric heating
11	0.96	Agriculture (plants)
12	0.96	Agriculture (plants) + electric sauna stove
13	0.96	Agriculture (dairy cattle)
14	0.96	Agriculture (dairy cattle) + electric sauna stove
15	0.96	Agriculture (dairy cattle) + electric heating + electric sauna stove
16	0.96	Meat production (big/chicken)
17	0.96	public service
18	0.96	private service
19	0.96	1-shift industry
20	0.96	2-shift industry
21	0.96	3-shift industry
22	0.96	street lights (clock)
23	0.96	street lights (pecu switch)
24	0.96	summer cottage
25	0.96	Markets
26	0.96	Hotels/restaurants
27	0.96	schools
28	0.96	Block of flats (service electricity)
29	0.96	hospital
30	0.96	heat-/waste management
31	0.96	YLE (radio broadcasting tower)
32	0.96	Inka (Factory)
33	0.96	Jita (Factory)
34	0.96	Kiilto (Factory)
35	0.96	Inhan tehtaat (Factory)
36	0.96	GWS (Factory)
37	0.96	Sports and culture
38	0.96	Scandic (hotel)