USING ADVANCED METERING INFRASTRUCTURE DATA FOR SMART GRID DEVELOPMENT

by

FRANKLIN L. QUILUMBA-GUDIÑO

Presented to the Faculty of the Graduate School of

The University of Texas at Arlington in Partial Fulfillment

of the Requirements

for the Degree of

DOCTOR OF PHILOSOPHY

THE UNIVERSITY OF TEXAS AT ARLINGTON

May 2014

Acknowledgements

This work research was funded by the DOE's Smart Grid Investment Grant for the Consolidated Edison Company of New York, Inc., and sponsorship of the author's graduate studies was provided by the Fulbright Commission, the Escuela Politécnica Nacional (Quito, Ecuador), and the Energy Systems Research Center at the University of Texas at Arlington.

February 12, 2014

Abstract

USING ADVANCED METERING INFRASTRUCTURE DATA

FOR SMART GRID DEVELOPMENT

Franklin L. Quilumba-Gudiño, PhD

The University of Texas at Arlington, 2014

Supervising Professor: Wei-Jen Lee

Identifying and using Advanced Metering Infrastructure (AMI) data to improve customer experience, utility operations, and advanced power management is one of the most important challenges in the smart grid development. Smart meters, capable of capturing frequent interval customer consumption (and possibly other parameters) using communication networks, are vital components of smart grid technology. Thus, smart meters expand the available range of data and functionality. Making the most of information from smart meters and smart grids increasingly requires dealing with Big Data. Big Data is a game changer, enabling utilities to transform the ways they interact with and serve their customers.

Today, many utilities are deploying smart meters as a vital step moving towards smart grids. Going from one meter reading per month to one meter reading at a sub-hourly rate (one minute, fifteen minutes, or thirty minutes) immediately poses a great technical challenge that can be overwhelming if not properly managed. AMI is becoming the standard in today's utility industry, making it possible to transform the performance of the grid and dramatically improve customer experience, utility operations, and advanced power management. To attain the maximum benefits from AMI, it is of utmost importance that utilities perform large-scale data analysis and transform them into information.

Consequently, this dissertation addresses the efforts involved in turning smart meter raw data into actionable information. Algorithms are developed to utilize data collected from AMI system for three main purposes:

1. To develop accurate customer daily load profiling for load estimation and network demand reconciliation to improve the efficiency and security of the utility grid.
2. To enhance the performance of load forecasting which impacts operating practices and planning decisions to build, lease, or sell generation and transmission assets and the decisions to purchase or sell power at wholesale level.
3. To investigate a nonintrusive load monitoring method for discerning individual appliances from a residential customer.

Table of Contents

List of Illustrations

List of Tables

Chapter 1

Introduction

1.1 Background

1.1.1 Smart Grid Initiative

In the recent years, significant progress has been achieved for advanced metering, specially supported by funding opportunities under the American Recovery and Reinvestment Act (ARRA). ARRA has placed a significant amount of funding in the hands of DOE, resulting in the Smart Grid Investment Grant (SGIG) program and the Smart Grid Demonstration (SGD) program ("Smart Grid Programs") [1]. The American Recovery and Reinvestment Act of 2009 is commonly referred to as "the Stimulus," and has three immediate goals:

- Create and save jobs
- Spur economic activity and invest in long-term growth by providing $288 billion in tax cuts and benefits; $224 billion to increase funding for entitlement programs; and $275 billion in contract, grant, and loan awards.
- Foster unprecedented levels of accountability and transparency in Recovery spending.

1.1.2 Smart Grid Programs

In accordance with ARRA main goals, to stimulate the economy and to create and save jobs [2], DOE launched these programs with orientation towards maximizing the public benefit, with particular interest in [1]:

1. Job Creation and Marketplace Innovation;
2. Peak Demand and Electricity Consumption;
3. Operational Efficiency;
4. Grid Reliability and Resilience;

5. Distributed Energy Resources and Renewable Energy; and
6. Carbon Dioxide Emissions.

1.1.3 Recovery Act - Smart Grid Investment Grant Program

The purpose of the Smart Grid Investment Grant Program (SGIG) is to accelerate the modernization of the nation's electric transmission and distribution systems and promote investments in smart grid technologies, tools, and techniques which increase flexibility, functionality, interoperability, cyber-security, situational awareness, and operational efficiency [3]. The purpose of this intended funding opportunity announcement is to stimulate the rapid deployment and integration of advanced digital technology that is needed to modernize the nation's electric delivery network for enhanced operational intelligence and connectivity. Applications are being sought that apply "smart" technology to: appliances and electrical equipment; electricity distribution and transmission systems; and homes, offices, and industrial facilities.

1.1.4 Advanced Metering Infrastructure Projects

Project applications in this topic area will be aimed at the installation of smart meters that can facilitate two-way communication between consumers and utilities. Smart meters are able to measure, store, send, and receive real time digital information concerning electricity use, costs, and prices that can be used to implement a range of customer service initiatives including dynamic pricing, demand response, load management, billing, remote connect/disconnect, outage detection and management, tamper detection, and other programs [4].

There are several awardees of this funding opportunity in different states with a total of $3.4 billion in investment to match a total public-private investment worth over $8 billion [5].

Under this program, Consolidated Edison Company of New York, Inc. has deployed a wide range of grid-related technologies, including automation, monitoring, and two-way communications to make the electric grid function more efficiently and enable the integration of renewable resources and energy efficient technologies. It will also benefit customers in New Jersey [5].

1.1.5 Advanced Metering Infrastructure

Advanced Metering Infrastructure (AMI) is an emerging technology evolving from Automated Meter Reading (AMR). Today, many utilities are deploying smart meters as a first step towards smart grids enabling the company and consumers to gather and utilize metered data in a more intelligent and cost effective manner.

The core role of revenue meters has always been to measure energy consumption in kilowatt hours (kWh) for billing purposes. Smart meters are no exception, but even more, smart meters vastly expand the available range of data and advanced functionality [6] to meet the evolving Smart Grid needs.

Electricity metering systems are varied in technology and design. Basically [7], smart meters collect data locally and transmit via a Local Area Network (LAN) to a data collector. This transmission can occur recurrently in 15-minute or hourly increments, or infrequently on a daily basis according to needs. The collector retrieves the data and may or may not carry out any processing of the data. Data is then transmitted via a Wide Area Network (WAN) to the utility central collection point for processing and use by business applications. Since the communications path is two-way, signals or commands can be sent directly to the meters. A basic architecture of Smart Meter System operations is shown in Figure 1-1.

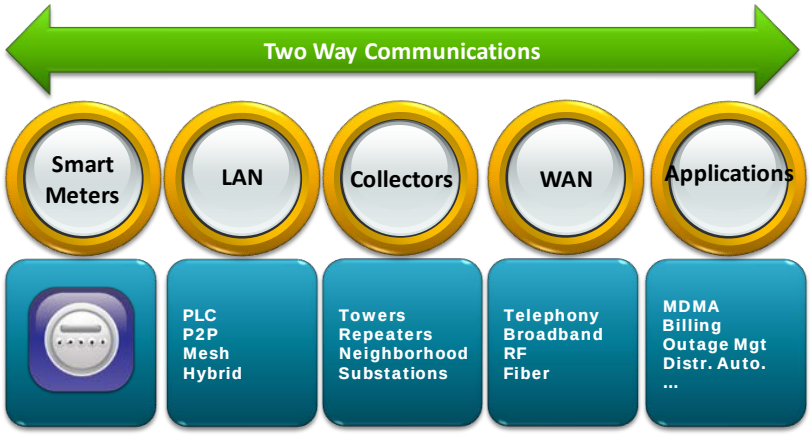

Figure 1-1 Advanced Metering Infrastructure Basic Architecture

Advanced Meters: Meters that measure and record usage data at hourly intervals or more frequently, and provide usage data to both consumers and energy companies at least once daily. Data are used for billing and other purposes. Advanced meters include basic hourly interval meters, meters with one-way communication, and real time meters with built-in two-way communication capable of recording and transmitting instantaneous data [8].

Some of the benefits of the deployment of AMI at the consumer and service provider level can be explained as follows [9]:

- Since smart meters communicate consumption data to both the user and the service provider, the consumers can be more aware of their energy usage with in-home displays. Going further, electric pricing information supplied by the service provider enables load control devices like smart thermostats to modulate electric demand based on pre-established consumer price preferences. More advanced customers deploy distributed energy resources (DER) based on these economic signals. Consumer portals process the AMI

data in ways that enable more intelligent energy consumption decisions, even providing interactive services like prepayment.

- On the other hand, the service provider (utility) can employ existing, enhanced or new back office systems that collect and analyze AMI data to help optimize operations, economics, and consumer service. For example, AMI provides immediate feedback on consumer outages and power quality, enabling the service provider to rapidly address grid deficiencies, and AMI's bidirectional communications infrastructure also supports grid automation at the station and circuit level. The vast amount of new data flowing from AMI allows improved management of utility assets as well as better planning of asset maintenance, additions, and replacements. The resulting more efficient and reliable grid is one of AMI's many benefits.

1.2 Motivation

The world is increasingly information-driven, and since smart meter deployments significantly increase data quantity and availability, data analytics becomes an essential piece of every electric utility company. The era of Big Data is here, but Big Data does not create value until it is transformed into useful information and is put into the context of solving important business challenges [10, 11]. To do so, it requires access to voluminous and varied data set, as well as strong analytics capabilities that include both software tools and the requisite skills to use them [12].

There are a number of challenges resulting from the greatly increased type, variability, volume, and timing of logging AMI data [13]. Data from multiple sources are not only generated in high volume, but they are also delivered at a high rate, outgrowing the ability for many traditional systems to store and analyze the data. As a result, much of the data is collected, but not analyzed [12].

Therefore, the motivation comes from this necessity of using the smart meter data to realize the benefits of the advanced metering infrastructure for smart grid development that gives utilities the ability to better manage their power grid and enables consumers to better control their consumption, among other benefits.

1.3 Contribution

Large amounts of available data inspire new ways to transform processes, organizations, entire industries, and even society itself. AMI is becoming the standard in today's utility industry, making it possible to transform the performance of the grid and dramatically improve customer experience, utility operations and advanced power management. To attain the maximum benefits from AMI, it is of utmost importance that utilities perform large-scale data analysis and transform them into information.

In data analysis, an analyst starts by preparing and preprocessing data used in the analytics tools. This task often receives little attention, and it is treated as a minor topic in the research literature and data-mining process [14]. In real world applications, the situation is reversed. Since limited efforts are focused on addressing load or customer power consumption data handling, the first contribution is to address AMI data preprocessing needed to turn raw data into actionable information.

Once the problems of AMI data have been carefully understood and solved during the data preprocessing stage, it is time to move forward.

Traditionally, utilities have collected and analyzed interval data for *a statistical sample of customers* of a particular type usually on a rate class basis [15]. Nowadays, with the deployment of smart meters, the availability of interval data is being extended to *all customers*.

In this sense, traditional statistical processes with sample statistics for estimation of class load profiles should be expanded. Therefore, the next contribution is to ease the

calculation of class load profiles by adding up across the interval data for customers in the class. Moreover, load profiles calculation is further enhanced by applying a well-established data mining technique, clustering, to identify patterns in load consumption.

Then, it is proposed that using clustering approaches applied to the load profiling problem will be an appealing idea for load forecasting because grouping load profiles based on consumption behavior similarities will reduce the variability of load which will be predicted over time, and therefore, the forecasting error.

Finally, an application of smart meter data for Non-Intrusive Load Monitoring (NILM) is proposed. A viable solution for a realistic development of NILM based on the AMI data is investigated considering polynomial load models as well.

1.4 Dissertation Outline

Before embarking on the task of advanced data analysis process, one must have a clear understanding of what kind of data the smart meters offer to generate valuable information. Real-world data are highly susceptible to noise, missing values, and inconsistency that must be resolved to be able to get the most out of the smart meter data. This is presented in Chapter 2.

Chapter 3 is dedicated to the development of load profiles. Model variables and algorithms are introduced for calculating load profiles based on stratification customer information, and also based on customers' behavior similarities.

Chapter 4 brings together the load profile clustering development to the load forecasting problem. This chapter presents how load forecasting at the system level can be further enhanced by applying a well-established data mining technique, clustering, to identify load consumption patterns with household level data from smart meters.

Inspired by non-intrusive load monitoring, in Chapter 5, a viable solution of using AMI data for Major Appliances Identification considering polynomial load models is investigated.

Finally, conclusions and idea of future works are presented in Chapter 6

Chapter 2

AMI Data Preprocessing

Large amount of data inspire new ways to transform processes, organizations, industries, and even society. Today, many utilities are deploying smart meters as the first step towards smart grids. Going from one meter reading per month to one meter reading every 15 minutes, or even every 1 minute immediately presents a great technical challenge that can be overwhelming if not properly managed. The process of identifying the pieces of AMI Data that contain value and determining how best to extract those pieces is critical. It may not be glamorous or exciting, but iteratively preprocessing AMI data, examining what it looks like, and adjusting the preprocess in order to better target the data that are needed is immensely important. Data preprocessing is a critical step to turn raw data into actionable information. Without completing this step, it will not be possible to proceed to the analysis phase.

Technical literature (many researchers) does not give the real importance to this matter because only the back-end is presented and the front-end is minimized due to assumptions that the data are ready to be used. Only a limited number of authors share their experiences on dealing with smart meter data, and the reason might be because smart meter data is probably beyond the reach of researchers outside the utility company, or only very limited data is being made available to consumers and researchers.

Real-world data are highly susceptible to noise, missing values, and inconsistency. Raw data that do not appear to show any of these problems should immediately arouse suspicion. Data quality is possibly the single most important factor to influence the quality of the results from any analysis. Despite the fact that data quality is a

subjective concept, data have quality if they satisfy the requirements of their intended use [16, 17].

2.1 Data Preparation

There are a number of challenges resulting from the greatly increased type, variability, volume, and timing of logging AMI data [13]. Data from multiple sources are not only generated in high volume, but are also delivered at a high rate, outgrowing the ability for many traditional systems to store and analyze the data. As a result, much of the data is collected, but not analyzed [12]. Moreover, because smart meters must transmit the near-real time data they gather to a central collection point, a lot of utilities struggle with the limited capacity of their communication networks. Understanding AMI data behavior and assessing AMI data quality are the beginning steps for AMI data analysis.

2.1.1 Familiarizing with AMI Data

Before embarking on the task of advanced data analysis process, one must have a clear understanding of what kind of data the smart meters offer and must be able to interpret the data. Without taking a closer look at attributes and data values, it will be difficult to generate valuable information even if the data are accurate, timely, consistent, and complete [18]. The core role of revenue meters has always been to measure energy consumption in kilowatt hours (kWh) for billing purposes, and smart meters are no exception. However, smart meters vastly expand the available range of data and advanced functionality [6] to meet the evolving Smart Grid needs.

2.1.1.1 Initial error checking

An obvious yet important first step in familiarizing with smart meter data is that one must be able to read the data even before beginning with any data check or any further steps in formal analysis [19].

In this step, one can easily identify that a set of ~1,500 smart meters can occupy up to ~2.5GB in disk as a common text format (.TXT) file for a single day.

2.1.1.2 Visualization

One of the simplest ways to gain insight on the data at hand would be visualization. It helps the analyst to become familiar with the data set. An appealing option is to plot the data over time: e.g. a day, a whole week, a whole month (Figure 2-1); or skim through the data set in its primary tabular form: on this way, one can distinguish attributes, data type (e.g. string, numeric), file delimiter, data resolution. Of course, only a very limited amount of data can be meaningfully presented and digested, but it may be enough to identify and establish a preliminary observation on the behavior of the raw data.

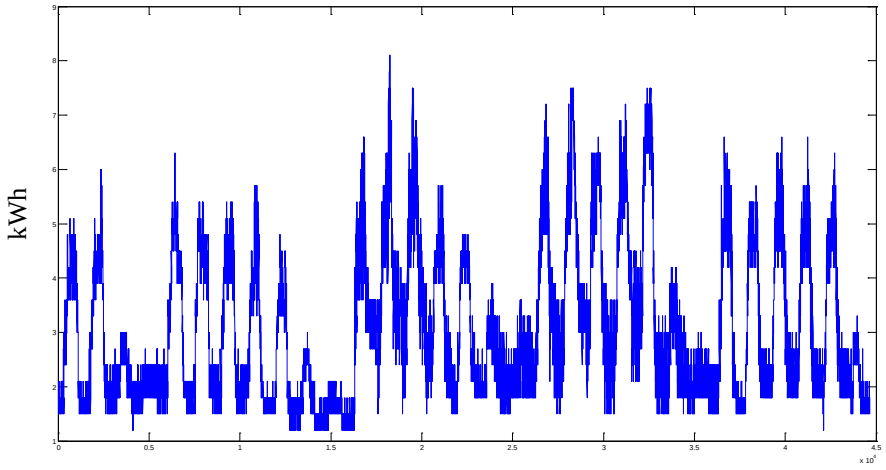

Figure 2-1 Channel 1 (kWh consumption), Time Interval: 1 minute, AMI-Data March 2012

2.1.1.3 Smart meter data format

In general, smart meter data are in the form of time series and are arranged in such a way that they come in the form of tuples. Starting from February 2012 until October 2013, three distinct smart meter data formats from a utility have been identified

due to system upgrades, as shown in Table 2-1. It should be pointed out that efforts have been made to establish a consistent format throughout the files starting from October 2012 and onwards.

Table 2-1 Smart Meter Data Format

Month	Available Attributes in Tuple Form
Feb-12	<Meter Number,Device_Mfg_Model,Channel,UOM,Raw Value,BlockEndValue,EndTime,Value,Revised_Value,CT_Ratio_value,Applied_CT_Ratio>
Mar-12	<Meter Number,Device_Mfg_Model,Channel,UOM,Raw Value,BlockEndValue,EndTime,Value,Revised_Value,CT_Ratio_value,Applied_CT_Ratio>
Apr-12	<Meter Number,Device_Mfg_Model,Channel,UOM,Raw Value,BlockEndValue,EndTime,Value,Revised_Value,CT_Ratio_value,Applied_CT_Ratio>
May-12	<MeterName,EndTime,Channel,RawValue,Value,UOM,BlockEndValue,Device_Mfg_Model>
Jun-12	<MeterName,EndTime,Channel,RawValue,Value,UOM,BlockEndValue,Device_Mfg_Model,Revised_Value,CT_Ratio_value,Applied_CT_Ratio>
Jul-12	<MeterName,EndTime,Channel,RawValue,Value,UOM,BlockEndValue,Device_Mfg_Model,Revised_Value,CT_Ratio_value,Applied_CT_Ratio>
Aug-12	<MeterName,EndTime,Channel,RawValue,Value,UOM,BlockEndValue,Device_Mfg_Model,Revised_Value,CT_Ratio_value,Applied_CT_Ratio>
Sep-12	<MeterName,EndTime,Channel,RawValue,Value,UOM,BlockEndValue,Device_Mfg_Model,Revised_Value,CT_Ratio_value,Applied_CT_Ratio>
Oct-12	<MeterName,EndTime,Channel,RawValue,Value,UOM,BlockEndValue,Device_Mfg_Model,Revised_Value,CT_Ratio_value,Applied_CT_Ratio>
…	…
Oct-13	<MeterName,EndTime,Channel,RawValue,Value,UOM,BlockEndValue,Device_Mfg_Model,Revised_Value,CT_Ratio_value,Applied_CT_Ratio>

Each of the distinct files is composed for the attributes shown in Table 2-2

Table 2-2 Attributes That Represent a Smart Meter Data

Attribute	Definition
Meter Number or Meter Name	Meter identification number
Channel	Measured quantity
Value	Actual measurement
End Time	Time stamp
UOM	Unit of measurement
Device_Mfg_Model	Device manufacturing model
CT_Ratio_value	Current transformer ratio

A typical single phase revenue meter for residential metering installations may store quantities or channels as shown in Table 2-3.

Table 2-3 Available Information From a Smart Meter

Channel #	Measurement	Operation
1	Total kWh	Sum
2	Total kVArh	Sum
3	Total Apparent kVAh	Sum
4	Current Phase A	Store
5	Current Phase B	Store
6	Current Phase C	Store
7	Voltage Line A to Neutral	Store
8	Voltage Line B to Neutral	Store
9	Voltage Line C to Neutral	Store
10	Current Phase A	Max
11	Current Phase B	Max
12	Current Phase C	Max
13	Voltage Line A to Neutral	Min
14	Voltage Line B to Neutral	Min
15	Voltage Line C to Neutral	Min
16	kWh Phase A	Sum
17	kWh Phase B	Sum
18	kWh Phase C	Sum
19	kVArh Phase A	Sum
20	kVArh Phase B	Sum

2.1.2 Smart Meter Data Resolution and Grouping

After becoming familiar with the AMI data at hand, and since data resolution or logging time is a concern, smart meters that record samples every 15 minutes and every 1 minute have been identified. Therefore, three very distinct smart meter files that depend on the recording time interval and the number of channels can be established:

Group 1. Min15Ch2 – Recordings every 15 minutes, 2 Channels,

Group 2. Min1Ch2 – Recordings every 1 minute, 2 Channels,

Group 3. Min1Ch20 – Recordings every 1 minute, 20 Channels.

2.2 Smart Meter Data Cleaning

As mentioned earlier, real-world data are highly susceptible to noise, missing values, and inconsistency. Data quality is possibly the single most important factor to influence the quality of the results from any analysis. This section discusses the encountered issues and the processes to resolve inconsistencies, deal with missing values, smooth noisy data, and identify or remove outliers.

2.2.1 Inconsistencies

When data processing centers make changes as a result of introducing improvements over time, downstream users may not be aware of these changes for a short interval of time. Some of these changes are obvious, but sometimes the changes are minor and difficult to ascertain [17]. The following are typical examples that were encountered:

- Changes in units of measurement. Meters present a wide flexibility on how they report their data. As shown in Figure 2-2, a meter was set to report in Wh initially and switched to kWh around 10 AM. A correction needs to be made to all affected meters.

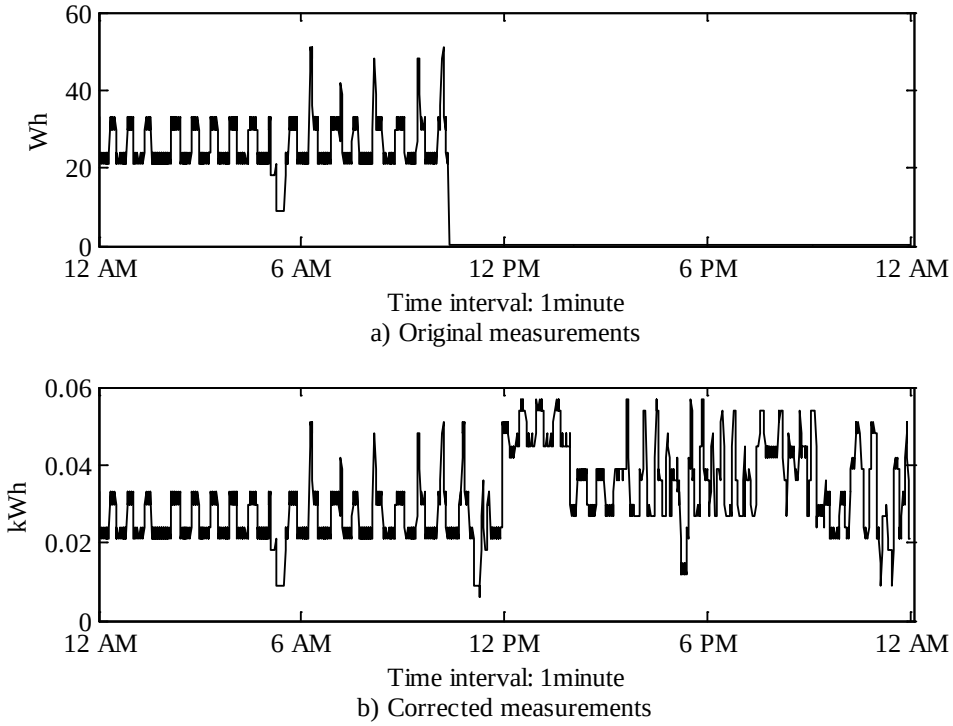

Figure 2-2 Changes in UOM Reporting in kWh Instead of Wh. If No Correction Was Made, It Would Be Mistakenly Thought That No Consumption Occurred.

- Swapping attributes for an entire data set. This can be noticed in Table 2-1 where the data provided at each column of the entire data set are effectively the same with the exception that, in some cases, CT ratio fields are not provided. A comparison of the relevant columns for this study is shown in Table 2-4, where Labels indicate the same data, and the Column indexes are matched, swapped (partially matched), or non-existent (partially matched). Partially matched means that a match occurs but not at all times.

Table 2-4 Comparison of Relevant Attributes Available on the Entire Data Set

Cases	Feb. 2012 - Apr. 2012		May 2012		Jun. 2012 and Onwards	
	Label	Col.	Label	Col.	Label	Col.
Matched	Meter Number	1	Meter Name	1	Meter Name	1
	Channel	3	Channel	3	Channel	3
Swapped (Partially Matched)	Value	8	Value	5	Value	5
	End Time	7	End Time	2	End Time	2
	UOM	4	UOM	6	UOM	6
	Device_Mfg_Model	2	Device_Mfg_Model	8	Device_Mfg_Model	8
Non-existent (Partially Matched)	CT_Ratio_value	10	CT_Ratio_value	NA	CT_Ratio_value	10

- Swapping rows for an entire data set. Generally, smart meter data sets contain their rows arranged by channel variables in ascending order, but data in May 2012 present row swapping. It was commonly found that rows were sorted in descending order in most of the smart meter files, for example in Table 2-5. However, some files have their rows swapped randomly, as shown in Table 2-6.

Table 2-5 Swapped Rows: Descending Order – Time Interval 15 Minutes, 2 Channels

MeterName	EndTime	**Channel**	RawValue	Value	UOM	BlockEndValue	Device_Mfg_Model
XXXXXX	2012-05-31 00:00:00	**2**	2	.0061000000	kVARh	1077.0000000000	I210+C
XXXXXX	2012-05-31 00:00:00	**1**	31	.0930000000	kWh	6957.0000000000	I210+C
XXXXXX	2012-05-31 00:15:00	**2**	0	.0000000000	kVARh	1077.0000000000	I210+C
XXXXXX	2012-05-31 00:15:00	**1**	22	.0660000000	kWh	6957.0000000000	I210+C
XXXXXX	2012-05-31 00:30:00	**2**	0	.0000000000	kVARh	1077.0000000000	I210+C
XXXXXX	2012-05-31 00:30:00	**1**	12	.0360000000	kWh	6958.0000000000	I210+C
…	…	…	…	…	…	…	…

Table 2-6 Swapped Rows: Random Order – Time Interval 1 Minute, 20 Channels

MeterName	EndTime	Channel	RawValue	Value	UOM	BlockEndValue	Device_Mfg_Model
XXXXXX	2012-05-19 00:00:00	20	0	.0000000000	kVARhB-N		kV2c
XXXXXX	2012-05-19 00:00:00	19	0	.0000000000	kVARhA-N	.0000000000	kV2c
XXXXXX	2012-05-19 00:00:00	18	1	.0030000000	kWhC-N	380.0000000000	kV2c
XXXXXX	2012-05-19 00:00:00	17	0	.0000000000	kWhB-N		kV2c
XXXXXX	2012-05-19 00:00:00	16	1	.0030000000	kWhA-N	169.0000000000	kV2c
XXXXXX	2012-05-19 00:00:00	15	1197	119.7000000000	VrmsC-N	117.0000000000	kV2c
XXXXXX	2012-05-19 00:00:00	14	0	.0000000000	VrmsB-N		kV2c
XXXXXX	2012-05-19 00:00:00	13	1201	120.1000000000	VrmsA-N	116.0000000000	kV2c
XXXXXX	2012-05-19 00:00:00	12	24	2.4000000000	IRMSC-N	16.0000000000	kV2c
XXXXXX	2012-05-19 00:00:00	11	0	.0000000000	IRMSB-N		kV2c
XXXXXX	2012-05-19 00:00:00	10	13	1.3000000000	IRMSA-N	2.0000000000	kV2c
XXXXXX	2012-05-19 00:00:00	9	1208	120.8000000000	VrmsC-N	121.0000000000	kV2c
XXXXXX	2012-05-19 00:00:00	8	0	.0000000000	VrmsB-N		kV2c
XXXXXX	2012-05-19 00:00:00	7	1209	120.9000000000	VrmsA-N	121.0000000000	kV2c
XXXXXX	2012-05-19 00:00:00	6	24	2.4000000000	IRMSC-N	2.0000000000	kV2c
XXXXXX	2012-05-19 00:00:00	5	0	.0000000000	IRMSB-N		kV2c
XXXXXX	2012-05-19 00:00:00	4	13	1.3000000000	IRMSA-N	1.0000000000	kV2c
XXXXXX	2012-05-19 00:00:00	3	3	.0090000000	kVAh	.0000000000	kV2c
XXXXXX	2012-05-19 00:00:00	2	0	.0000000000	kVARh	.0000000000	kV2c
XXXXXX	2012-05-19 00:00:00	1	1	.0030000000	kWh	549.0000000000	kV2c
…	…	…	…	…	…	…	…
XXXXXX	2012-05-19 16:43:00	12	23	2.3000000000	IRMSC-N	16.0000000000	kV2c
XXXXXX	2012-05-19 16:43:00	11	0	.0000000000	IRMSB-N		kV2c
XXXXXX	2012-05-19 16:44:00	1	1	.0030000000	kWh	554.0000000000	kV2c
XXXXXX	2012-05-19 16:44:00	3	2	.0060000000	kVAh	.0000000000	kV2c
XXXXXX	2012-05-19 16:44:00	2	0	.0000000000	kVARh	.0000000000	kV2c
XXXXXX	2012-05-19 16:44:00	7	1204	120.4000000000	VrmsA-N	121.0000000000	kV2c
XXXXXX	2012-05-19 16:44:00	6	23	2.3000000000	IRMSC-N	2.0000000000	kV2c
XXXXXX	2012-05-19 16:44:00	5	0	.0000000000	IRMSB-N		kV2c
XXXXXX	2012-05-19 16:44:00	4	13	1.3000000000	IRMSA-N	1.0000000000	kV2c
XXXXXX	2012-05-19 16:44:00	20	0	.0000000000	kVARhB-N		kV2c
XXXXXX	2012-05-19 16:44:00	19	0	.0000000000	kVARhA-N	.0000000000	kV2c
XXXXXX	2012-05-19 16:44:00	18	1	.0030000000	kWhC-N	384.0000000000	kV2c
XXXXXX	2012-05-19 16:44:00	17	0	.0000000000	kWhB-N		kV2c
XXXXXX	2012-05-19 16:44:00	16	1	.0030000000	kWhA-N	170.0000000000	kV2c
XXXXXX	2012-05-19 16:44:00	15	1198	119.8000000000	VrmsC-N	117.0000000000	kV2c
XXXXXX	2012-05-19 16:44:00	14	0	.0000000000	VrmsB-N		kV2c
XXXXXX	2012-05-19 16:44:00	13	1203	120.3000000000	VrmsA-N	116.0000000000	kV2c
XXXXXX	2012-05-19 16:44:00	12	24	2.4000000000	IRMSC-N	16.0000000000	kV2c
XXXXXX	2012-05-19 16:44:00	11	0	.0000000000	IRMSB-N		kV2c
XXXXXX	2012-05-19 16:44:00	10	13	1.3000000000	IRMSA-N	2.0000000000	kV2c
XXXXXX	2012-05-19 16:44:00	9	1199	119.9000000000	VrmsC-N	120.0000000000	kV2c
XXXXXX	2012-05-19 16:44:00	8	0	.0000000000	VrmsB-N		kV2c
…	…	…	…	…	…	…	…

- Getting more or less attributes. As mentioned earlier, utility personnel has the option to turn on or off the data channels from smart meter. This provides flexibility in information gathering but greatly increases the complexity of data conversion process.
- Changes in format. In some situations an attribute or a variable is expressed in a different way. For example, the attribute “Meter Number” is sometimes labeled as “MeterName” (Table 2-4), which represents the same data attribute. Another example can be seen in Table 2-7 regarding the Time Stamp.

Table 2-7 Changes in Format for Timestamp Data

Timestamp Format	Characteristics
‘[m]m/[d]d/yyyy [H]H:MM:SS XX’ *Examples:* *2/8/2012 10:39:00 PM* *3/11/2012 9:45:00 AM* *4/3/2012 8:02:00 PM* *11/11/2012 2:24:00 AM*	Values in brackets are optional 12-hour clock format XX could be either AM or PM Date separator is ‘/’ Time separator is ‘:’ Date-Time separator is ‘ ’
‘yyyy-mm-dd HH:MM:SS’ *Examples:* *2012-05-09 00:03:00* *2012-06-04 13:44:00* *2012-07-10 16:15:00* *2012-08-12 08:09:00* *2012-09-17 20:03:00*	24-hour clock format Date separator is ‘-’ Time separator is ‘:’ Date-Time separator is ‘ ’

As solutions to inconsistency problems, in addition to propagate and cycle notifications whenever a change is made anywhere in the system to ensure that every entity which uses that data element is informed [17], a comprehensive software design should take these situations into consideration.

2.2.2 Missing Data

Incomplete data are commonplace properties of large real-world databases and data warehouses, and it can occur for a number of reasons. Some attribute values are not recorded because they are considered irrelevant [16]. The data collection instruments used may be faulty. Similarly, errors in data transmission, if either the receiving end or the sending end have problems [16, 17]. In addition, the entering of day light saving can also cause missing data.

The simplest solution for this problem is the reduction of the data set by eliminating all samples with missing values. This is possible when large data sets are available [14]. Fill-in methods such as replacing a missing attribute value by a measure of central tendency for the attribute (e.g. mean or median) or assigning the most probable value to the missing attribute could be considered [20]. Figure 2-3 shows the missing data due to the beginning of the daylight saving time (DST).

According to the experiences in dealing with smart meter data, two distinct types of missing data were found: missing interval data and missing channel data. Missing interval data could be associated with transmission problems or day light saving while missing channel data occurs because no data have been stored in that channel. These cases have to be treated differently when processing the data.

2.2.3 Duplicate Data

Communication interruption and resend requests may create duplicate data. Generally, data cleaning can be performed to detect and remove redundancies in the data [16]. By solving row data inconsistencies, duplicate data can be handled. The other possibility of data duplication is the end of day light saving. They have to be identified and treated as valid data. Figure 2-4 shows duplicate data due to the ending of day light saving.

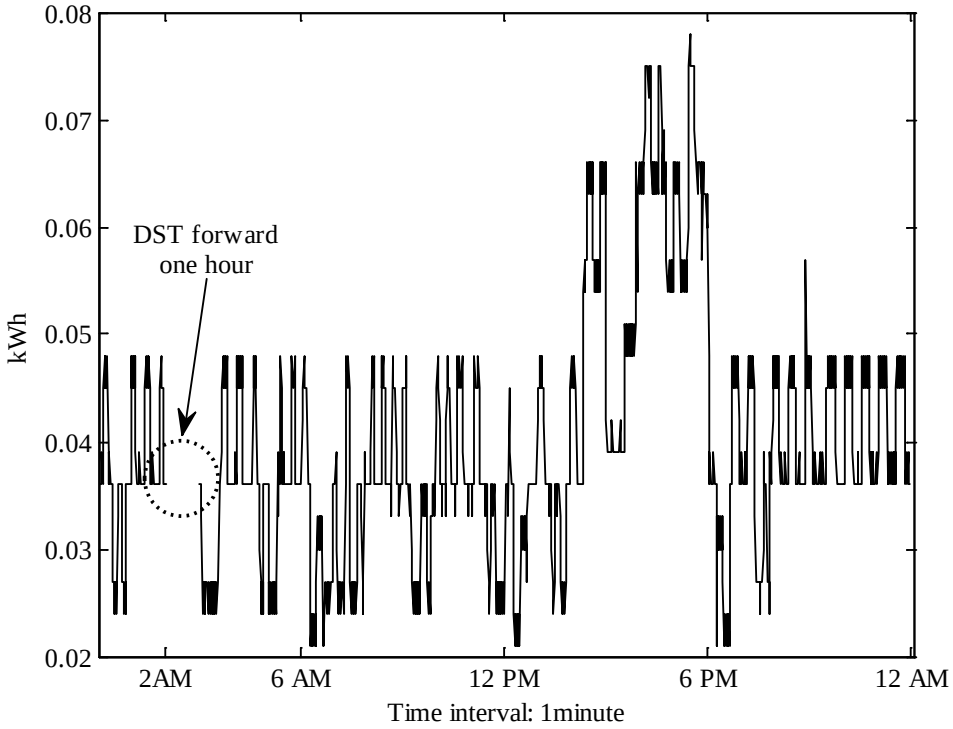

Figure 2-3 “Missing Data” Due to Daylight Saving Time in the United States That Began at 2:00 AM on Sunday, March 11, 2012

2.2.4 Outlier Detection

Alternatively, data cleaning tasks include outlier detection. Samples that are significantly different or inconsistent with the remaining set of data are called outliers. Outlier detection and clustering analysis are two highly related tasks [16], where a cluster of small sizes can be considered as clustered outliers [21].

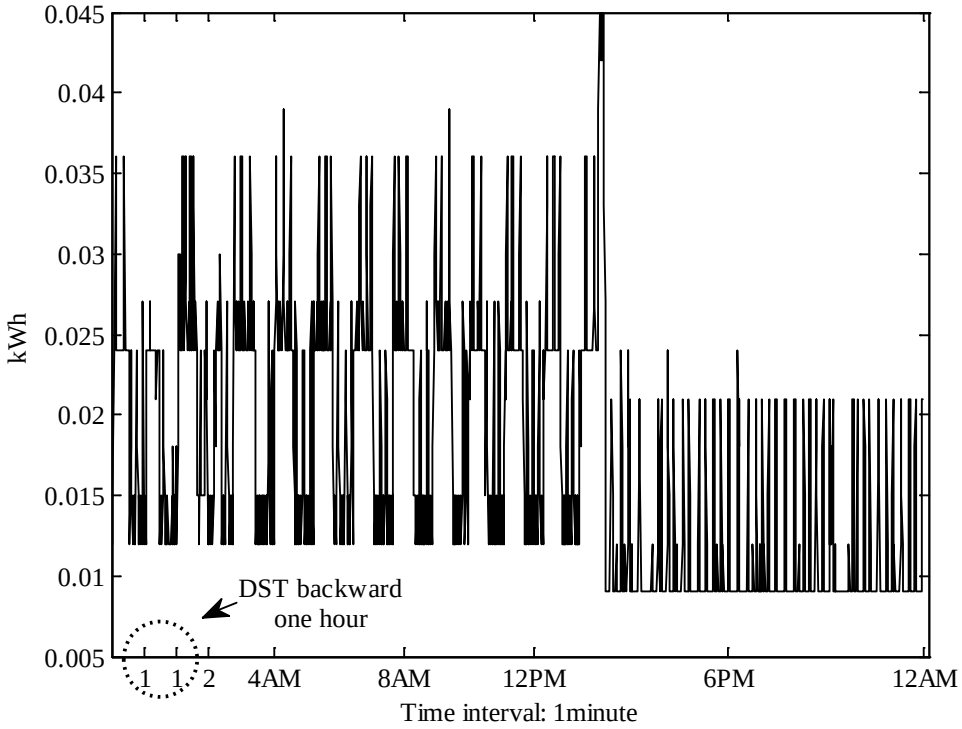

Figure 2-4 “Duplicate Data” Due to Daylight Saving Time in the United States That Ended at 2:00 AM on Sunday, November 4, 2012

2.3 Data Preprocessing Software Design Criteria

Data reading and preprocessing software is designed to be robust enough to deal with and solve the issues that have been introduced. Some are fixable while others are not, but the ultimate goal is to prepare the data for analysis.

First of all, the raw files should be in the known format as it has been presented in Table 2-1. These are large files containing all the meters at a particular day, so to be able to manage these data, each line of a large file is read, and a new file is created every time a new smart meter is found. After this step is performed, a single file per meter, per day at each month is obtained, as shown in Figure 2-5.

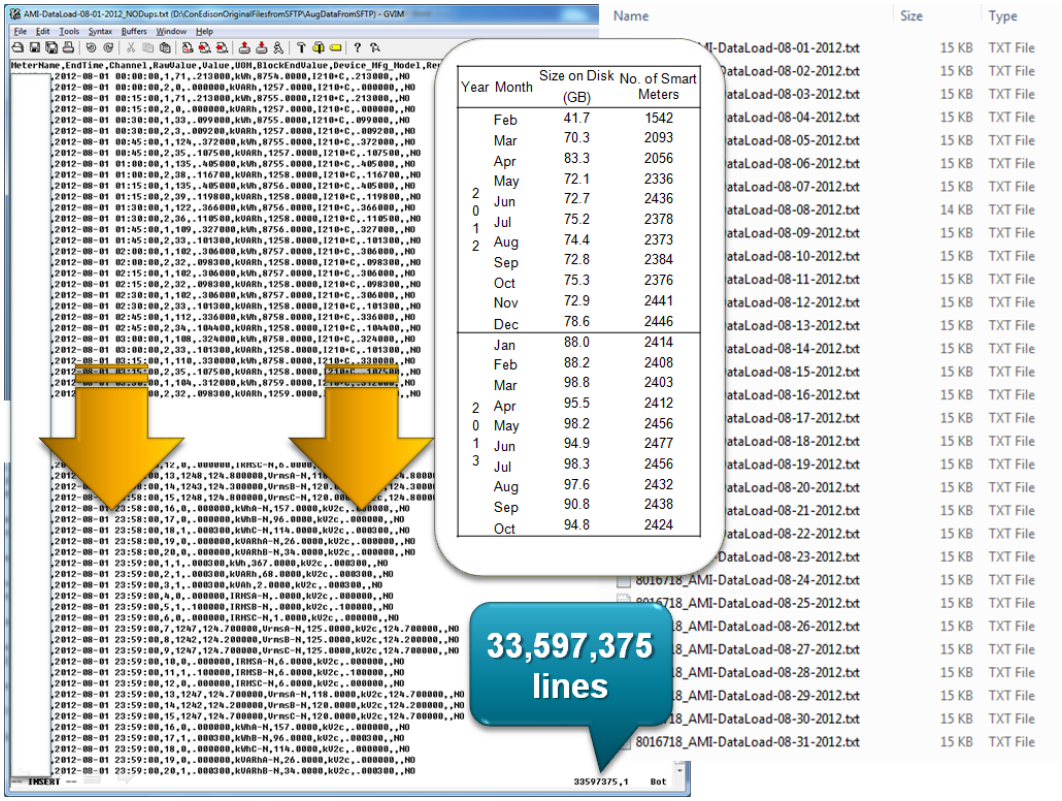

Figure 2-5 A Single Text File Obtained From the AMI System Can Contain Up to 33 Million Lines of Data. This File Corresponds To 1-Day Worth of Data on August 1, 2012, and It Has Up to 2373 Smart Meters. Each Smart Meter is Then Allocated in a Single File per Meter, Per Day to Be More Manageable. One May Notice That Each Month a Great Amount of Data Is Collected Under This New Era of Smart Meters.

In addition, a Raw Count is collected which indicates the number of data lines that each meter has at a particular day, this helps in identifying whether the large files contain reasonable data. Then, a data validation step is performed where not only inconsistencies and missing/duplicate values are taken care of but also each smart meter file is converted into a binary file containing its data for the whole month to make it convenient for handling. Moreover, the data file format is standardized, making it possible

to have smart meter files with the same meaning, attributes, and fields; but with a different number of channels and time interval, as shown in Table 2-8.

Table 2-8 Standard Format for the Smart Meter Data Files

Attribute	Definition
Meter Number	Unique smart meter identification number
Date	Date with the following format '*YYYYMMDD'*
Data	A matrix of *NoOfPointsPerDay*-by-*NoOfChannels* ▪ The *NoOfPointsPerDay* will be 96 for a resolution of 15 minutes, or 1440 for a resolution of 1 minute. ▪ The *NoOfChannels* will be 2 or 20.

Finally, each smart meter file is organized in a group: Min15Ch2, Min1Ch2, and Min1Ch20. In addition, a Detailed Count is kept for each smart meter at each day of the month per channel to be utilized by utility personnel, and Detailed Indicators explain the problems the raw data has for each smart meter at each day the whole month per Channel. As examples of this, summary counts for Channel 1 at each group of smart meters are provided in Table 2-9, Table 2-10, and Table 2-11.

Table 2-9 Summary Count from February 2012 to October 2013 – Min15Ch2-Channel 1

Year						2012					
Month	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
Total No. of Meters	593	593	599	598	598	595	597	597	596	596	593
No. of Days	29	31	30	31	30	31	31	30	31	30	31
No. of Meters with < 4 days missing	592	593	597	596	595	595	584	594	594	593	588
Percentage	99.83	100.00	99.67	99.67	99.50	100.00	97.82	99.50	99.66	99.50	99.16
Year						2013					
Month	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	
Total No. of Meters	588	589	582	582	597	596	593	588	586	587	
No. of Days	31	28	31	30	31	30	31	31	30	31	
No. of Meters with < 4 days missing	584	577	580	581	580	591	585	583	585	585	
Percentage	99.32	97.96	99.66	99.83	97.15	99.16	98.65	99.15	99.83	99.66	

Table 2-10 Summary Count from February 2012 to October 2013 – Min1Ch2-Channel 1

Year	2012										
Month	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
Total No. of Meters	250	256	265	266	261	258	258	258	261	263	264
No. of Days	29	31	30	31	30	31	31	30	31	30	31
No. of Meters with < 4 days missing	237	250	256	261	255	258	244	258	258	259	262
Percentage	94.80	97.66	96.60	98.12	97.70	100.00	94.57	100.00	98.85	98.48	99.24
Year	2013										
Month	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	
Total No. of Meters	262	245	241	241	269	269	265	257	263	262	
No. of Days	31	28	31	30	31	30	31	31	30	31	
No. of Meters with < 4 days missing	239	232	241	240	239	257	256	254	255	256	
Percentage	91.22	94.69	100.00	99.59	88.85	95.54	96.60	98.83	96.96	97.71	

Table 2-11 Summary Count from February 2012 to October 2013 – Min1Ch20-Channel 1

Year	2012										
Month	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
Total No. of Meters	699	1244	1192	1472	1577	1525	1518	1529	1519	1582	1589
No. of Days	29	31	30	31	30	31	31	30	31	30	31
No. of Meters with < 4 days missing	492	665	1136	1074	1335	1412	665	1153	1315	1128	1447
Percentage	70.39	53.46	95.30	72.96	84.65	92.59	43.81	75.41	86.57	71.30	91.06
Year	2013										
Month	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	
Total No. of Meters	1564	1574	1580	1589	1590	1612	1598	1587	1589	1575	
No. of Days	31	28	31	30	31	30	31	31	30	31	
No. of Meters with < 4 days missing	1375	1482	1425	1499	1399	1334	1222	1245	1099	1159	
Percentage	87.92	94.16	90.19	94.34	87.99	82.75	76.47	78.45	69.16	73.59	

It should be pointed out that the Detailed Count is analyzed by channel. Depending on the application, one or more channels are revised to determine valid smart meter files for subsequent use if desired.

Chapter 3

AMI Data for Load Profiling

Load profile estimation has been identified as one of the most desirable applications after AMI implementation since the load shapes, as well as the daily peak load, are vital factors in scheduling, operation, and control of the utility grid. In the long term, demand forecasting is useful to plan and purchase power supply by utilities, schedule equipment maintenance, and provide an early warning to consumers of potential load curtailment or advanced pricing information. In the short term, it is essential to know with as much accuracy as possible what the total and local system demand will be in the next minutes, hours, and days so that generators with different costs and constraints can be scheduled to optimize total system efficiency.

Hourly load profiles have been in use for a long time. Typically, electric service companies estimate the load of consumers on a class-by-class basis, using smooth, 24-hour peak day load curves. Customers are linked to one of the predefined classes, and the load of each customer is then estimated with customer class-specific hourly load profiles. Each household within the same class will have identical normalized daily load curve, although each will be different because each home has different appliances and is occupied by people with different schedules and usage preferences. Typically, only active power is considered for calculating Load Profiles, and there is no distinction for three or single phase consumption. Load curves have been on a 15, 30, 60, and 120 minute basis.

Based upon the following two criteria, this dissertation presents two approaches to develop load profiles for residential customers:

- The customer stratification information
- The customers' behavior similarities

3.1 Model Variables

The groups of variables that appear in these models are:

Data Meters Variables

- Recording Time Interval
    - 15-minutes smart meter data
    - 1-minute smart meter data
- Service Class and Stratification

Calendar Variables

- Day of the week
- Holiday
- Weekday and weekend
- Season of the year

Each of these groups of variables is discussed separately in the following sections.

3.1.1 Data Meters Variables

The profile data that are used as the dependent variable in the profile models are developed from meter readings for individual customers from the Utility AMI system. There are essentially three very distinct groups of meters based on their data logging capabilities. One group of the utility data sets provided individual data at the 15-minute level. Although, the remaining data sets were at the 1-minute level, the load profiles were developed at 15-minute basis as required by the utility personnel. Along with the data sets, information on the Service Class and Stratification Groups for the underlying customers was provided.

The Utility tariff establishes the Service Classes definitions. Service Class 1 is considered for the present study, and it corresponds to Residential Customers. Stratum

Category represents a subgroup within a service class. It is a measure of the size of a customer as defined by a particular billing quantity. Table 3-1 relates the stratum variable to the studied Service Class 1.

Table 3-1 Service Class 1 and its Stratum Billing Variable

Service Class	Service Class Description	Stratum Billing Variable
SC 1	Residential (excluding Religious and Water Heating)	Annual kWh

There are six stratum groups within service class 1 – residential. Table 3-2 shows the annual kWh ranges for the stratification variables. Group A includes those customers with no usage; these may be unoccupied apartments. Group F has no upper limit; these may be quite large.

Table 3-2 Service Class 1 and its Stratum Billing Variable

# Class	Group	Stratification Criteria	Low (kWh)	High (kWh)
RESID	A	Total Annual kWh	0	1948
RESID	B	Total Annual kWh	1949	2897
RESID	C	Total Annual kWh	2898	3897
RESID	D	Total Annual kWh	3898	5239
RESID	E	Total Annual kWh	5240	7741
RESID	F	Total Annual kWh	7742	9999999999

Therefore, to characterize the residential customer load curves, Service Class and Stratum Group should be known a priori for all the meters that are subject to analysis for developing the load profiles based on stratification information, but not when the load profiles are developed based on behavior similarities, where the customers groups are inferred from the data.

3.1.2 Calendar Variables

The main calendar variables include the day of the week, indicators of season, and holiday schedules.

3.1.2.1 Day of the week variables

The variables used in the models are shown in Table 3-3.

Table 3-3 Day of the Week Variables

Day of the week index	Days
1	Sundays
2	Mondays
3	Tuesdays
4	Wednesdays
5	Thursdays
6	Fridays
7	Saturdays

These variables are used when the load profiles are to be developed for a Day Type. The following provides a discussion of the importance of these variables:

- Saturday and Sunday. Residential loads tend to be slightly higher than on weekdays, reflecting the fact that most people are home from work and school.
- Monday. Monday loads tend to be slightly different from days in the middle of the week. People coming from a short break during weekend tend to show a different behavior, and starting again their activities at work or school.
- Tuesday, Wednesday, and Thursday. These days in the middle of the week tend to be highly similar in residential loads.
- Friday. Friday loads tend to be slightly different from days in the middle of the week. Residential loads extend into later hours.

3.1.2.2 Holiday variables

Specific dates are introduced for each individual holiday. Weekday holidays have higher residential loads than typical weekdays. The exact effect on the loads depends on the holiday. The following is a list (Table 3-4) of all specific holidays that are included in the Utility Schedule. The date has been expressed as an integer number to facilitate its processing.

Table 3-4 Holiday Day-Type Schedule 2012 - 2013

Holiday	Date	Integer Date
New Year's Day	2-Jan-12	20120102
Presidents' Day	20-Feb-12	20120220
Memorial Day	28-May-12	20120528
Independence Day	4-Jul-12	20120704
Labor Day	3-Sep-12	20120903
Thanksgiving Day	22-Nov-12	20121122
Christmas Day	25-Dec-12	20121225
New Year's Day	1-Jan-13	20130101
Presidents' Day	18-Feb-13	20130218
Memorial Day	27-May-13	20130527
Independence Day	4-Jul-13	20130704
Labor Day	2-Sep-13	20130902
Thanksgiving Day	28-Nov-13	20131128
Christmas Day	25-Dec-13	20131225

3.1.2.3 Weekday and weekend variables

In addition to Day of the Week variables, to give flexibility to the Utility Company when creating load profiles, Weekday and Weekend Variables are specified to create the load profiles for weekdays and weekends only without days of the week distinction.

Weekend and holiday electrical demand are presumed to exhibit behaviors different from the demand behavior of typical weekdays. A "Weekday" corresponds to

any weekday that is not a major holiday. Likewise, “Weekend” is defined to be the complement of the Weekday variable.

3.1.2.4 Season of the year variables

Residential and commercial loads typically exhibit strong seasonal variation arising from operation of heating, ventilation, and air conditioning loads. It then follows that four season variables are defined that are applicable to the specific Utility Company under study.

- Summer for summer months: June, July, August, and September
- Fall for fall months: October
- Winter for winter months: November, December, January, and February
- Spring for spring months: March, April, and May

3.2 Load Profile Development Based on Stratification Customer Information

The objective of this part is to develop valid and useable load profiles. Because smart meter is at an endpoint, it can be aggregated in different ways to serve load profile purposes. This has a significant implication because the load profile development follows a straightforward, structural approach based on Service Class and Stratification information. Service Class represents a group of customer types with similar load characteristics, and Stratum Category represents a subgroup within a service class.

Therefore, the smart meters are grouped by stratum categories, and the time period can be defined by the user ranging from a day, a month, or a season. For instance, a day could be determined to be at system peak demand, a month could be any of the available months, and a season could be summer.

As required for load profile development, curves are defined at 15-minute intervals, resulting in a 96-point daily curve. In general, for load profiling only Channel 1 is needed, which corresponds to total kWh. Most of the smart meters show their channel 1

with not many problems as data integrity is concerned, containing full data most of the time.

When more than one day (month or season) is considered, a meter should be present at all time. Then, the first step is to average load consumption at each individual meter for the whole time period. Two cases are considered:

1. Load profile considering only Stratification Data without Day Type distinction except Weekday, Weekend, or Holiday if specified.
2. Load Profile considering only Stratification Data at each Day Type of the Week: Holidays, Mondays, Tuesdays, Wednesdays, Thursdays, Fridays, Saturdays and Sundays.

Once individual load consumption at the meter level is determined, the next step is to proceed with a load profile for consumers that belong to a stratum category. The load profile is characterized by the mean and standard deviation curves. Due to the large amount of load profile cases that can be obtained from the current data, presenting exhaustive results would not be convenient.

As an example of the construction of load profiles based on the stratification information, load profiles for summer corresponding to the months of June, July, August, and September of 2012 are calculated. The smart meters belong to the group of Min15Ch2. There are 541 smart meters in total, where 100 meters belong to Stratum A, 131 meters to Stratum B, 205 meters to Stratum C, 83 meters to Stratum D, 22 meters to Stratum E, and no meters belong to Stratum F. Figure 3-1 to Figure 3-10 show the Load Profiles for Weekdays and Weekends during Summer 2012.

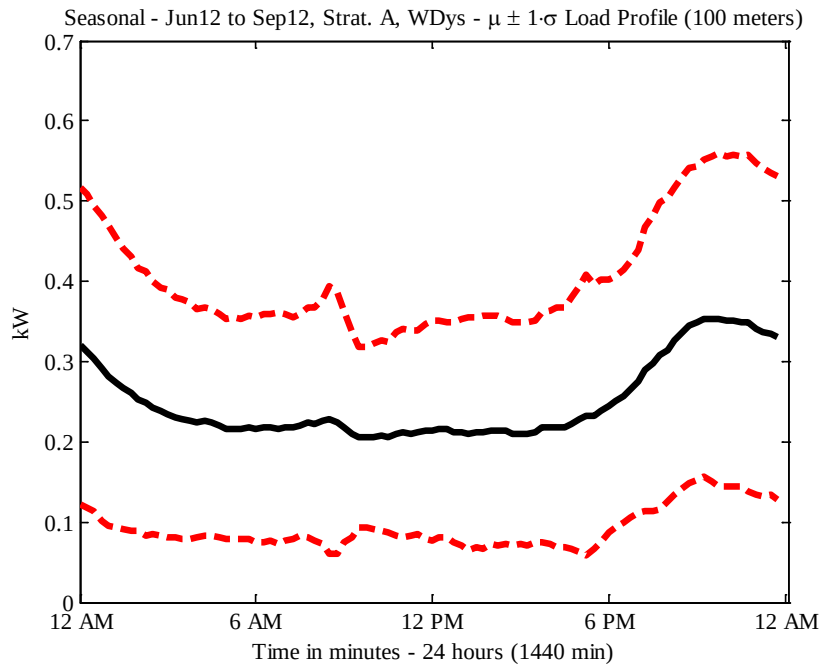

Figure 3-1 Weekdays Load Profile – Mean ± Std. Dev. for Stratum A

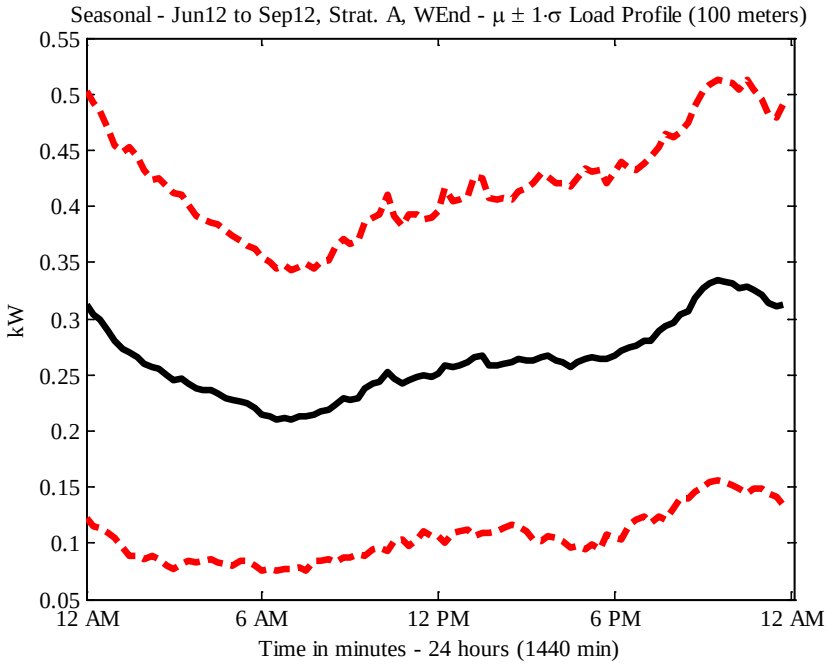

Figure 3-2 Weekends Load Profile – Mean ± Std. Dev. for Stratum A

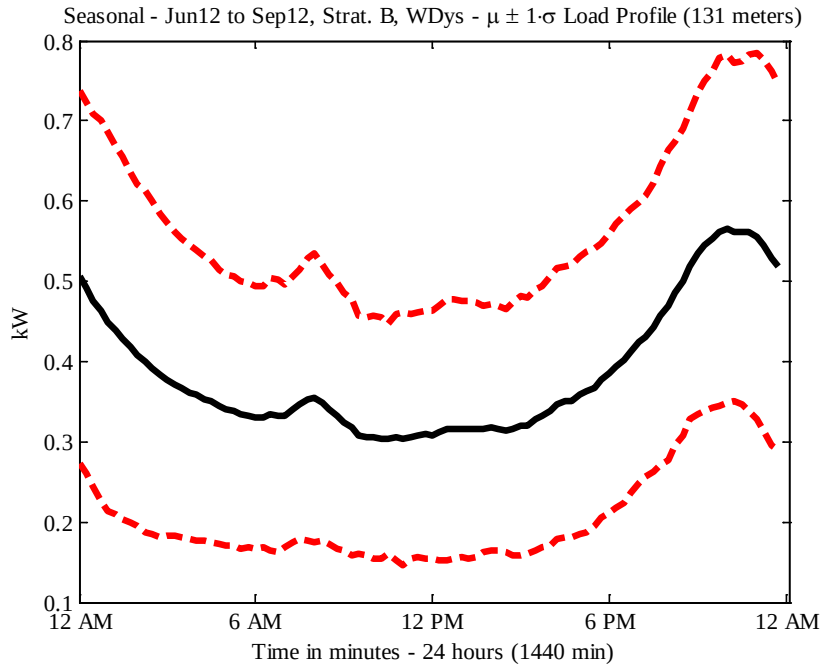

Figure 3-3 Weekdays Load Profile – Mean ± Std. Dev. for Stratum B

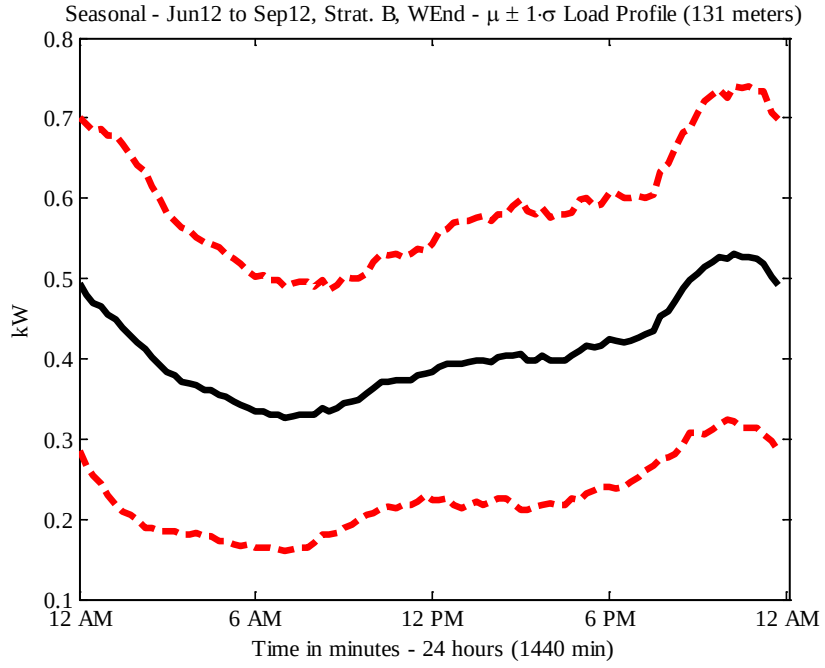

Figure 3-4 Weekends Load Profile – Mean ± Std. Dev. for Stratum B

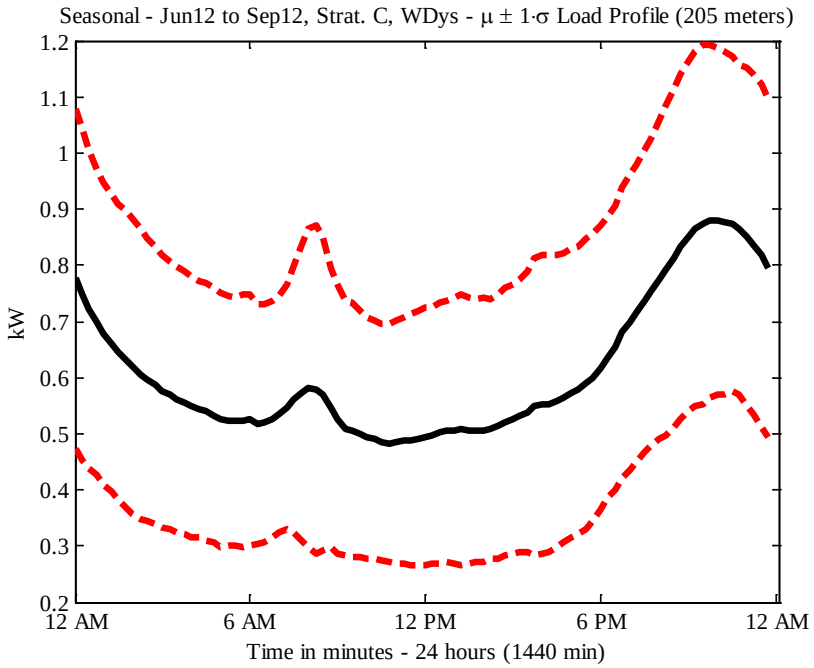

Figure 3-5 Weekdays Load Profile – Mean ± Std. Dev. for Stratum C

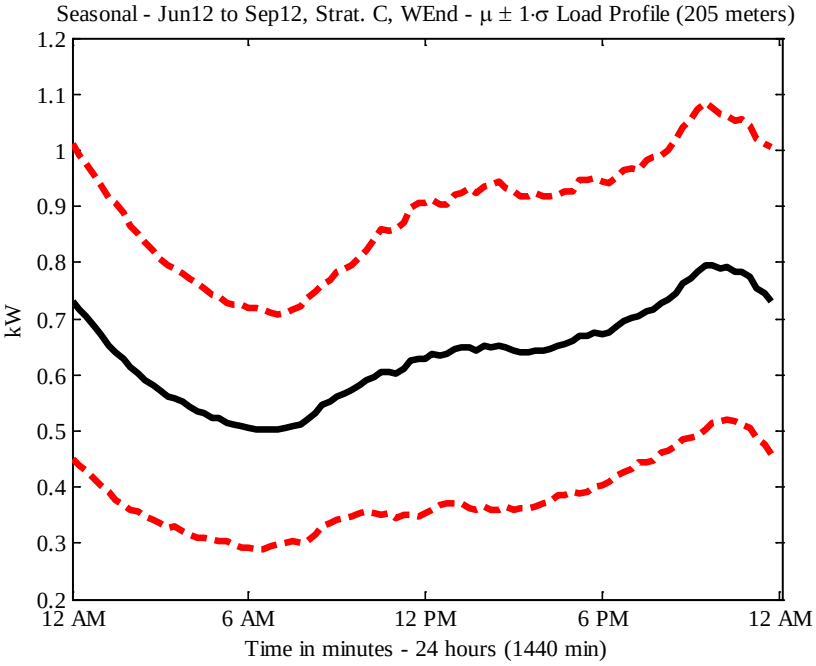

Figure 3-6 Weekends Load Profile – Mean ± Std. Dev. for Stratum C

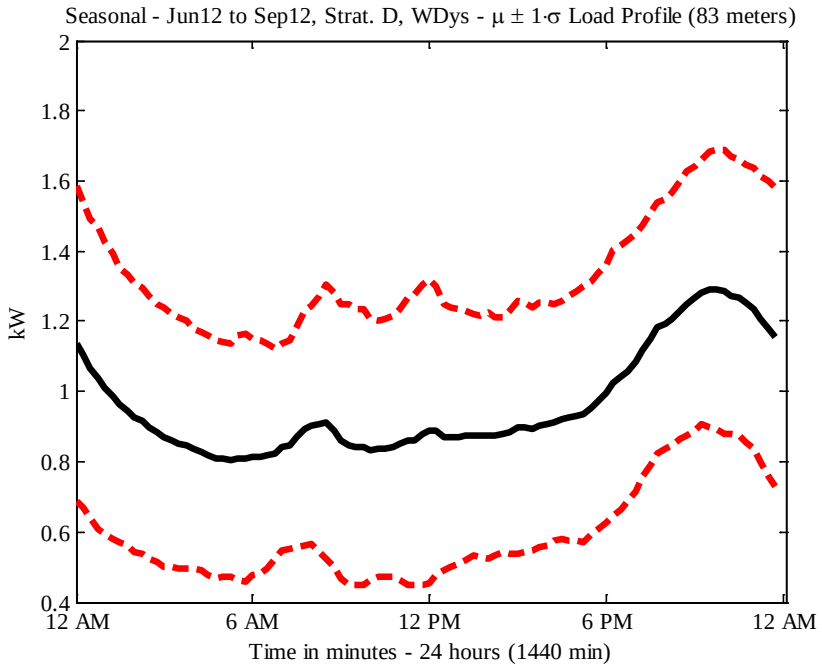

Figure 3-7 Weekdays Load Profile – Mean ± Std. Dev. for Stratum D

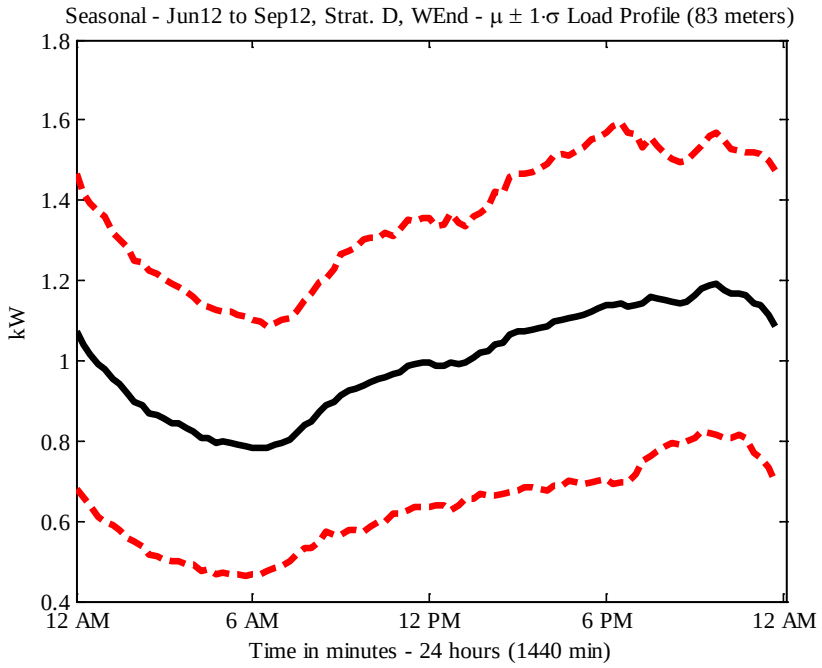

Figure 3-8 Weekends Load Profile – Mean ± Std. Dev. for Stratum D

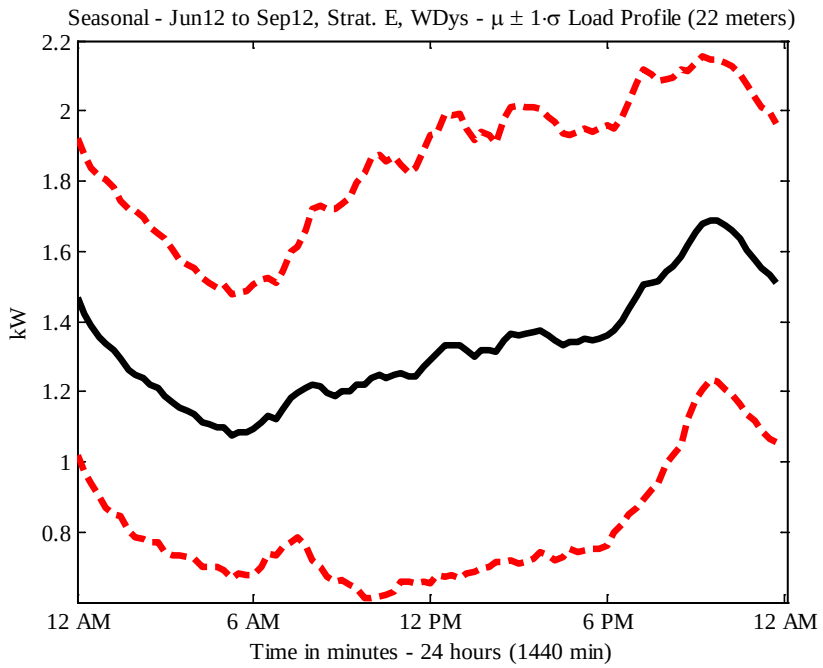

Figure 3-9 Weekdays Load Profile – Mean ± Std. Dev. for Stratum E

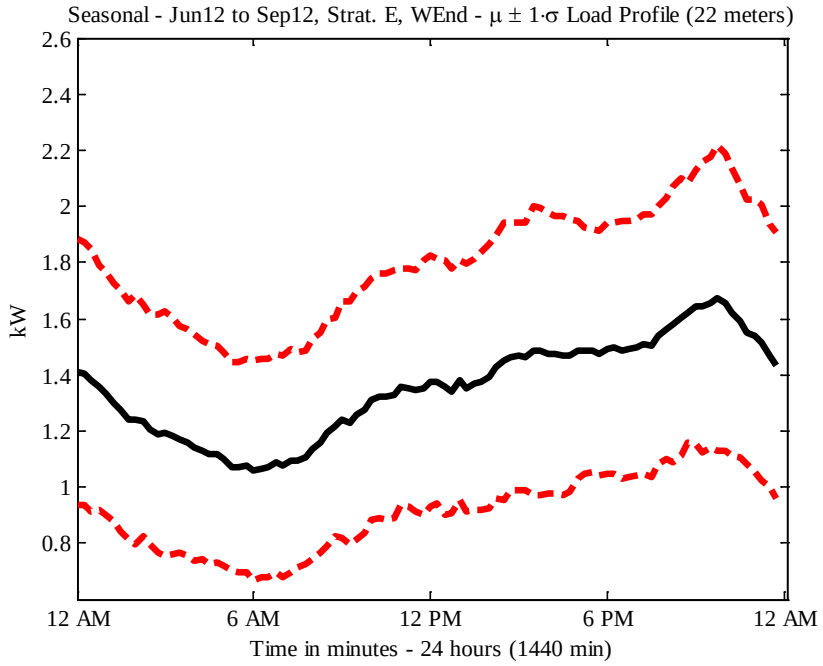

Figure 3-10 Weekends Load Profile – Mean ± Std. Dev. for Stratum E

To provide examples of Load Profiles for Day Type considering stratification information, only Stratum C is plotted for the sake of clarity from Figure 3-11 to Figure 3-18. Nevertheless, load profiles can be calculated for all stratum categories.

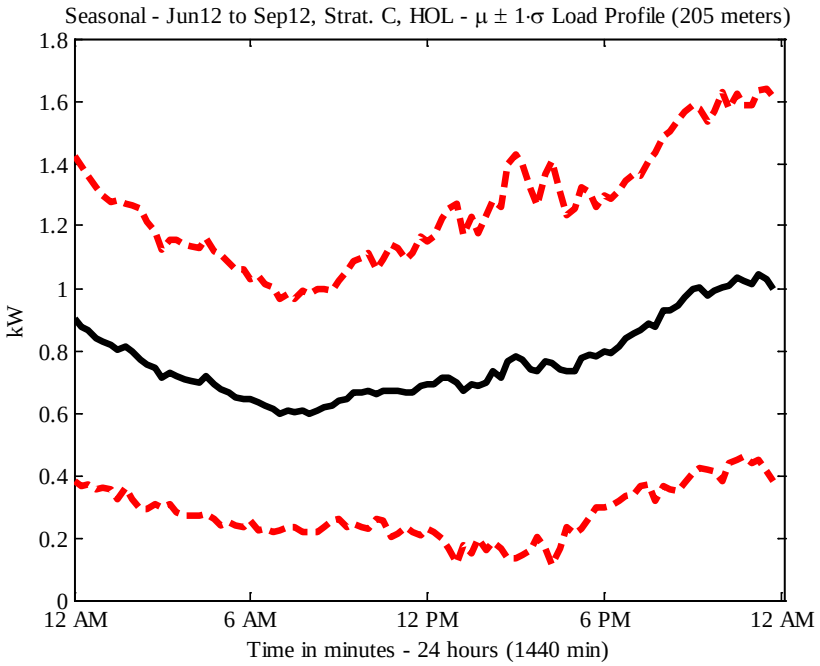

Figure 3-11 Holidays Load Profile – Mean ± Std. Dev. for Stratum C

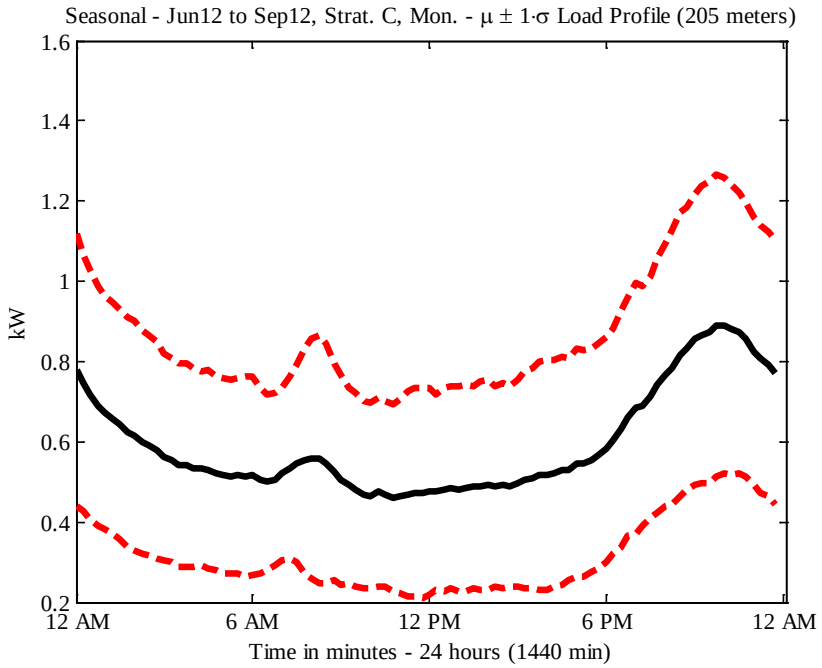

Figure 3-12 Mondays Load Profile – Mean ± Std. Dev. for Stratum C

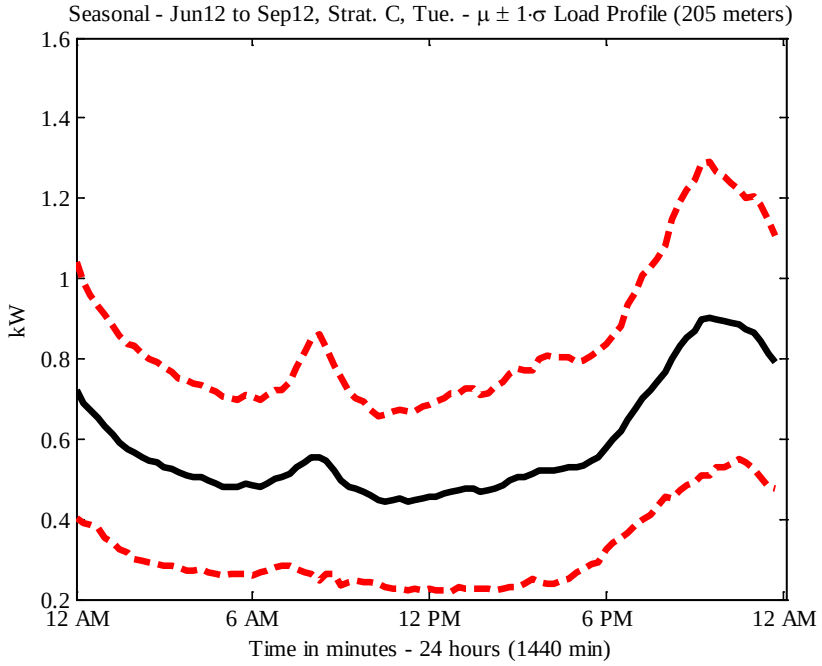

Figure 3-13 Tuesdays Load Profile – Mean ± Std. Dev. for Stratum C

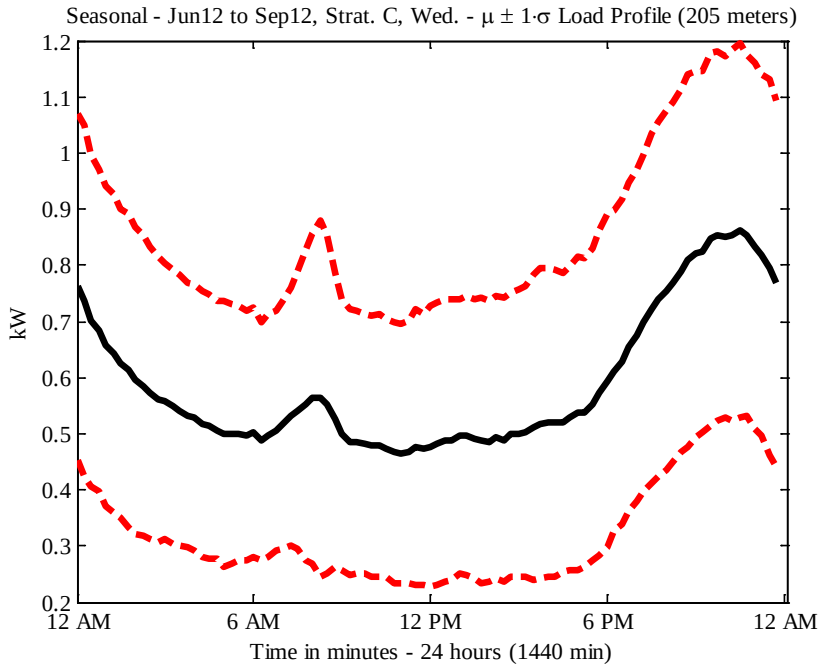

Figure 3-14 Wednesdays Load Profile – Mean ± Std. Dev. for Stratum C

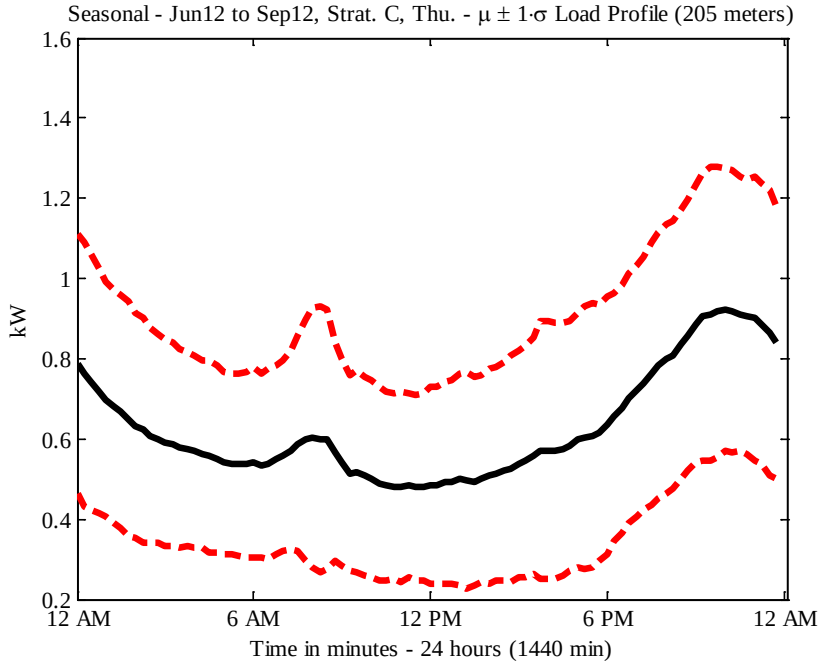

Figure 3-15 Thursdays Load Profile – Mean ± Std. Dev. for Stratum C

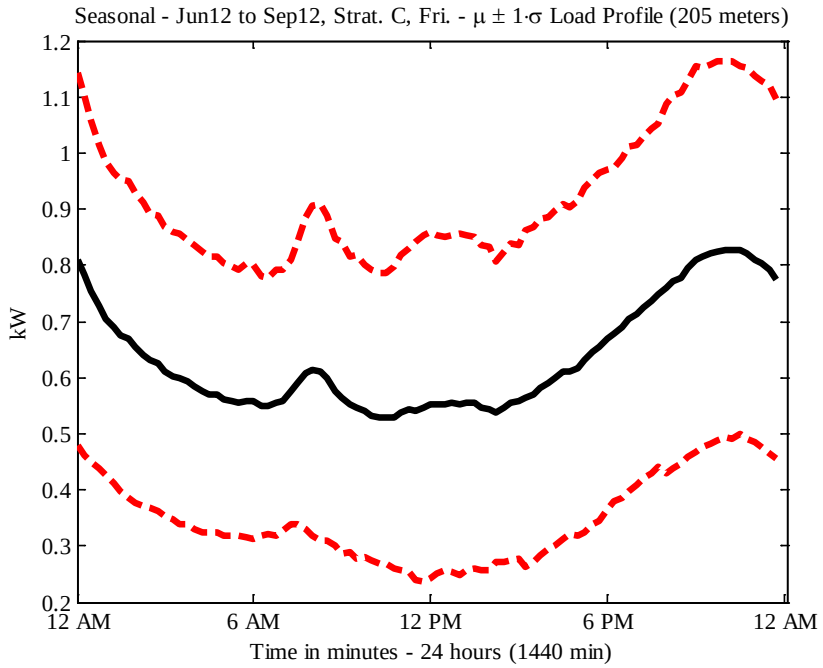

Figure 3-16 Fridays Load Profile – Mean ± Std. Dev. for Stratum C

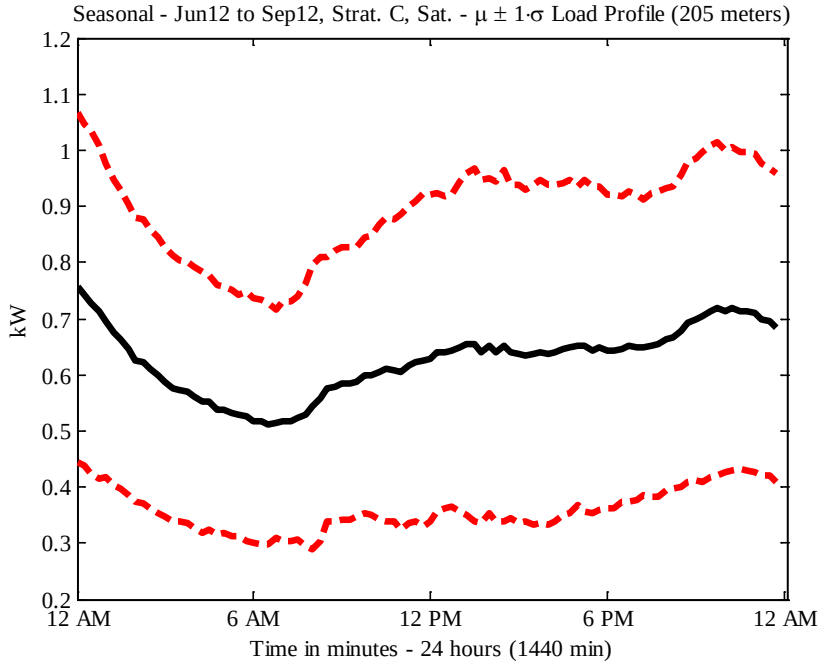

Figure 3-17 Saturdays Load Profile – Mean ± Std. Dev. for Stratum C

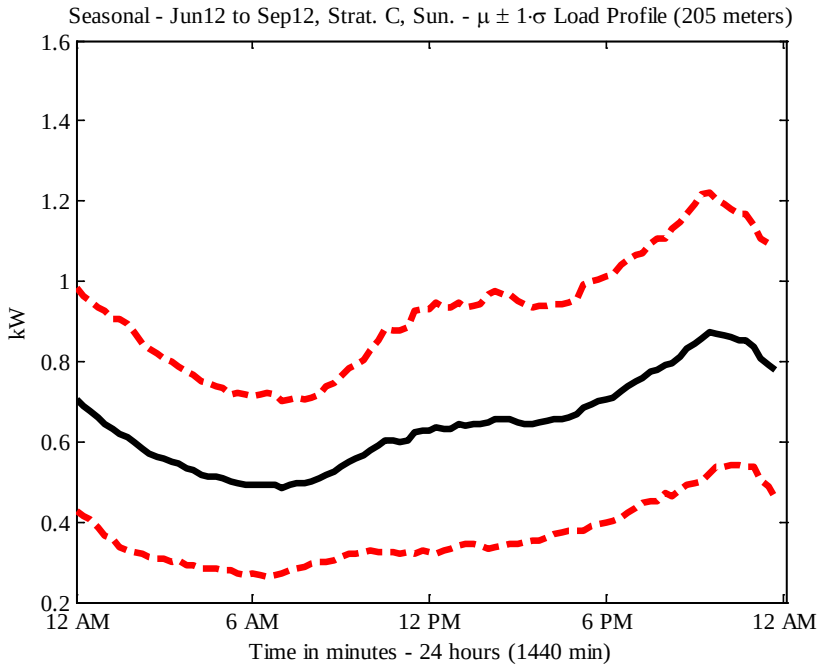

Figure 3-18 Sundays Load Profile – Mean ± Std. Dev. for Stratum C

3.3 Load Profile Development Based on Customers' Behavior Similarities

The purpose of this part of the dissertation is to develop load profiles of residential customers based on customers' behavior similarities by applying a well-established data mining technique: clustering. Clustering has been applied in a wide variety of fields [23]: engineering (computational intelligence, machine learning, pattern recognition, mechanical engineering, and electrical engineering), computer sciences (web mining, spatial database analysis, information retrieval, textual document collection, and image segmentation), life and medical sciences (genetics, biology, microbiology, paleontology, psychiatry, clinic, phylogeny, pathology), astronomy and earth sciences (geography, geology, remote sensing), social sciences (sociology, psychology, archeology, anthropology, education), and economics (marketing, business).

First of all, data clustering will be introduced, and then the application of clustering to obtain the load profiles will be presented.

3.3.1 Introduction to Data Clustering

Data clustering[1] is concerned with exploring data sets to assess whether or not they can be summarized meaningfully in terms of a relatively small number of groups or clusters [24]. The goal of clustering is to discover the natural grouping(s) of a set of patterns, points, objects, or individuals. Given a representation of n objects, the goal is to find K groups based on a measure of similarity such that the similarities between objects in the same group are high while the similarities between objects in different groups are low [25].An ideal cluster can be defined as a set of points that is compact and isolated.

Generally, clustering problems can be divided into two categories (Figure 3-19) [26]: hard clustering (or crisp clustering) and fuzzy clustering (or soft clustering). In hard clustering, a data point belongs to only one cluster; while in fuzzy clustering, a data point may belong to two or more clusters with some probabilities. A fuzzy clustering can be converted to a hard clustering by assigning each pattern to the cluster with the largest measure of membership [27]. Conventional clustering algorithms can be divided into two categories: hierarchical and partitional. Hierarchical clustering algorithms recursively find nested clusters either in agglomerative mode (starting with each data point in its own cluster and merging the most similar pair of clusters successively to form a cluster hierarchy) or in divisive (top-down) mode (starting with all the data points in one cluster and recursively dividing each cluster into smaller clusters). Compared to hierarchical clustering algorithms, partitional clustering algorithms find all the clusters simultaneously as a partition of the data and do not impose a hierarchical structure [25].

[1] Data clustering (or just clustering), also called cluster analysis, segmentation analysis, taxonomy analysis, or unsupervised classification

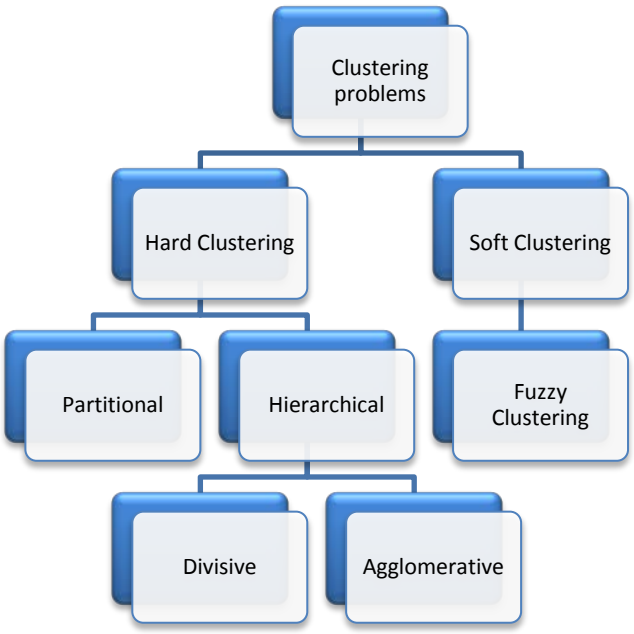

Figure 3-19 Diagram of Clustering Algorithms

3.3.2 Clustering Definition

Mathematically, a data set[2] with m objects, each of which is described by n attributes[3], is denoted by $D = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \dots, \boldsymbol{x}_i, \dots, \boldsymbol{x}_m\}$, where $\boldsymbol{x}_i = (x_{i1}, x_{i2}, \dots, x_{in})$, with each scalar measure x_{ij} denoting the jth component or attribute of $\boldsymbol{x}_i$. This data set to be clustered is viewed as an $m \times n$ pattern matrix. Each row in the matrix denotes an object while each column represents a feature [23, 27].

3.3.2.1 Proximity measures

Clustering algorithms are defined over sets of entities at which a proximity measure has been or can be defined. Proximity is the generalization of both dissimilarity and similarity [23]. Dissimilarities or distance functions are typically non-negative real numbers: the closer entities are to each other the smaller the dissimilarities are,

[2] 'Data set' or 'pattern set'
[3] Attribute, feature, dimension or variable

decreasing to zero to express the identity case. In contrast, similarities can be negative and they are increased to express closer ties between entities [28].

Given the data set D, each object of which is described by a *d*-dimensional feature vector, the distance matrix for D is defined as [26],

$$M_{dist}(D) = \begin{pmatrix} 0 & d_{12} & \dots & d_{1m} \\ d_{21} & 0 & \dots & d_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ d_{m1} & d_{m2} & \dots & 0 \end{pmatrix} \qquad (3.1)$$

where $d_{ij} = d(\boldsymbol{x}_i, \boldsymbol{x}_j)$ with respect to some distance function $d(\cdot,\cdot)$.

The similarity matrix for D is defined as [57],

$$M_{sim}(D) = \begin{pmatrix} 1 & s_{12} & \dots & s_{1m} \\ s_{21} & 1 & \dots & s_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ s_{m1} & s_{m2} & \dots & 1 \end{pmatrix} \qquad (3.2)$$

where $s_{ij} = s(\boldsymbol{x}_i, \boldsymbol{x}_j)$ with respect to some similarity function $d(\cdot,\cdot)$.

3.3.2.1.1 Proximity measures for continuous variables

It is most common to calculate the dissimilarity between patterns using distance measure defined on the features space. A distance measure such as those shown in Table 3-5 is chosen to evaluate the dissimilarity between any two clusters centroids, or feature vectors.

Consider two data points $\boldsymbol{x}_i = (x_{i1}, x_{i2}, \dots, x_{in})$ and $\boldsymbol{x}_j = (x_{j1}, x_{j2}, \dots, x_{jn})$, for example. The Euclidean distance is calculated as

$$d(\boldsymbol{x}_i, \boldsymbol{x}_j) = \left(\sum_{k=1}^{n} (x_{ik} - x_{jk})^2\right)^{\frac{1}{2}} \qquad (3.3)$$

Table 3-5 Dissimilarity Measures for Computing Distances

Distance measure	Equation
Euclidean	$d_{euc}(\boldsymbol{x}_i, \boldsymbol{x}_j) = \left(\sum_{k=1}^{n} (x_{ik} - x_{jk})^2\right)^{\frac{1}{2}} = \left[(\boldsymbol{x}_i - \boldsymbol{x}_j)(\boldsymbol{x}_i - \boldsymbol{x}_j)^T\right]^{\frac{1}{2}}$
Squared Euc.	$d_{seuc}(\boldsymbol{x}_i, \boldsymbol{x}_j) = \sum_{k=1}^{n} (x_{ik} - x_{jk})^2 = (\boldsymbol{x}_i - \boldsymbol{x}_j)(\boldsymbol{x}_i - \boldsymbol{x}_j)^T$
Manhattan or City block	$d_{man}(\boldsymbol{x}_i, \boldsymbol{x}_j) = \sum_{k=1}^{n} \lvert x_{ik} - x_{jk} \rvert$
Minkowski	$d_{mink}(\boldsymbol{x}_i, \boldsymbol{x}_j) = \left(\sum_{k=1}^{n} \lvert x_{ik} - x_{jk} \rvert^p\right)^{\frac{1}{p}}, p \geq 1$
Mahalanobis	$d_{mah}(\boldsymbol{x}_i, \boldsymbol{x}_j) = \left[(\boldsymbol{x}_i - \boldsymbol{x}_j)\Sigma^{-1}(\boldsymbol{x}_i - \boldsymbol{x}_j)^T\right]^{\frac{1}{2}}$
Canberra	$d_{can}(\boldsymbol{x}_i, \boldsymbol{x}_j) = \sum_{k=1}^{n} \frac{\lvert x_{ik} - x_{jk} \rvert}{\lvert x_{ik} \rvert + \lvert x_{jk} \rvert}$
Chebychev	$d_{che}(\boldsymbol{x}_i, \boldsymbol{x}_j) = \max_{1 \leq k \leq n} \lvert x_{ik} - x_{jk} \rvert$
Cosine	$d_{cos}(\boldsymbol{x}_i, \boldsymbol{x}_j) = 1 - \frac{\sum_{k=1}^{n} x_{ik} x_{jk}}{\left(\sum_{k=1}^{n} x_{ik}^2 \sum_{k=1}^{n} x_{jk}^2\right)^{\frac{1}{2}}}$

n: number of attributes; x_{ik}: attribute *k* of feature vector $\boldsymbol{x}_i$ in cluster-1; x_{jk}: attribute *k* of feature vector $\boldsymbol{x}_j$ in cluster-2; ^{T}: transpose of matrix; Σ: covariance matrix; *p*: order of Minkowski distance.

3.3.2.2 Clustering algorithms

Despite the fact that there are a large number of clustering algorithms, there is no correct answer on which one is the best, as it highly depends on the nature of the dataset and what constitutes meaningful clusters in an application [25].

3.3.2.2.1 Hierarchical clustering

In hierarchical clustering, the data are not partitioned into a particular cluster in a single step. Instead, a series of partitions takes place, which may run from a single cluster containing all objects to *k* clusters with each containing a single object. Hierarchical clustering is subdivided into *agglomerative* methods, which proceed by a

series of successive fusions of the *n* individuals into groups, and *divisive* methods, which separate the *n* individuals successively into finer groupings [24, 29]. Once divisions or fusions are made, the data-points assigned to a cluster cannot move to another cluster. Agglomerative techniques are more commonly used.

Hierarchical Agglomerative Clustering Algorithm [27]

1. Compute the proximity matrix containing the distance between each pair of patterns. Treat each pattern as a cluster.
2. Find the most similar pair of clusters using the proximity matrix. Merge these two clusters into one cluster. Update the proximity matrix to reflect this merge operation.
3. If all patterns are in one cluster, stop. Otherwise, go to step 2.

Based on the way the proximity matrix is updated in step 2, a variety of agglomerative algorithms can be designed. There are four common options [30]:

Single linkage. The distance between two clusters is the distance between the two closest data points in these clusters (each point taken from a different cluster).

Complete linkage. The distance between two clusters is the distance between the two furthest data points in these clusters.

Average linkage. Both single linkage and complete linkage are sensitive to outliers. Average linkage provides an improvement by defining the distance between two clusters as the average of the distances between all pairs of points in the two clusters.

Ward's method. At each step of agglomerative clustering, instead of merging the two clusters that minimize the pairwise distance between clusters, Ward's method merges the two clusters that minimize the 'information loss' for the step. The 'information loss' is measured by the change in the sum of the squared error of the clusters before

and after the merge. In this way, Ward's method assesses the quality of the merged cluster at each step or the agglomerative procedure.

3.3.2.2.2 Partitional clustering

Partitional methods are very efficient in applications involving large data sets. Partitional techniques usually produce clusters by optimizing a criterion function defined either locally (on a subset of the patterns) or globally (defined over all of the patterns) [27]. One of the most commonly used optimization-based methods is k-means clustering. In this algorithm, the number of clusters *k* is assumed to be fixed. There is an error function in this algorithm. It proceeds, for a given initial *k* clusters, by allocating the remaining data to the nearest clusters and then repeatedly changing the membership of the clusters according to the error function until the error function does not change significantly or the membership of the clusters no longer changes [26].

k-Means Clustering Algorithm

Let D be a data set with *n* instances, and let $C_1, C_2, \dots, C_k$ be the *k* disjoint clusters of D. Then the error function is defined as

$$E = \sum_{i=1}^{k} \sum_{\boldsymbol{x} \in C_i} d\big(\boldsymbol{x}, \mu(C_i)\big) \qquad (3.4)$$

where $\mu(C_i)$ is the centroid of cluster C_i ; $d\big(\boldsymbol{x}, \mu(C_i)\big)$ denotes the distance between $\boldsymbol{x}$ and $\mu(C_i)$, and it can be one of the many distance measures discussed before, a typical choice of which is the Euclidean distance $d_{euc}(\cdot,\cdot)$ as defined in Table 3-5.

1. Choose k cluster centers to coincide with k randomly-chosen patterns or k randomly defined points inside the hypervolume containing the pattern set.
2. Assign each pattern to the closest cluster center.
3. Re-compute the cluster centers using the current cluster memberships.

4. If a convergence criterion is not met, go to step 2. Typical convergence criteria are: no (or minimal) reassignment of patterns to new cluster centers, or minimal decrease in squared error.

Typically, k-means is run independently for different values of *k* and the partition that appears the most meaningful to the domain expert is selected. Different initializations can lead to different final clustering because k-means only converges to local minima. One way to overcome the local minima is to run the k-means algorithm, for a given *k*, with multiple different initial partitions and choose the partition with the smallest squared error [25].

3.3.2.2.3 Fuzzy clustering

So far hard (or crisp) clustering algorithms require that each data point of the data set belong to one and only one cluster. Fuzzy clustering extends this notion to associate each data point in the data set with every cluster using a membership function [26]. This gives the flexibility to express that data points can belong to more than one cluster [31].

Let D be a data set with *n* objects, each of which is described by *d* attributes, and let *c* be an integer between one and *n*. Then a fuzzy c-partition is defined by a c × n matrix $U = (u_{li})$ that satisfies

$$u_{li} \in [0,1], \qquad 1 \le l \le c,\ 1 \le i \le n\ , \qquad (3.5)$$

$$\sum_{l=1}^{c} u_{li} = 1\,, \qquad 1 \le i \le n\ , \qquad (3.6)$$

$$\sum_{l=1}^{n} u_{li} > 0\,, \qquad 1 \le l \le c\ , \qquad (3.7)$$

where u_{li} denotes the degree of membership of the object i in the lth cluster.

For each fuzzy c-partition, there is a corresponding hard c-partition. Let u_{li} $(l = 1,2,\dots,c, i = 1,2,\dots,n)$ be the membership of any fuzzy c-partition. Then the corresponding hard c-partition of u_{li}can be defined as ω_{li} as follows [26]:

$$\omega_{li} = \begin{cases} 1, & if \ \ l = \arg\ \max_{1 \le j \le c} u_{ji} \\ 0, & otherwise \end{cases} \qquad (3.8)$$

3.3.2.2.4 Affinity propagation clustering

Affinity propagation (AP) takes as input measures of similarity between pairs of data points and simultaneously considers all data points as potential exemplars[4]. Real-valued messages are exchanged between data points until a high-quality set of exemplars and corresponding clusters gradually emerges [32]. Affinity propagation aims at maximizing the net similarity, where clusters gradually emerge during the message-passing procedure [33].

Affinity propagation algorithm

Let s be the similarity matrix with a collection of real-valued similarities between data points, where the similarity $s(i,k)$ indicates how well the data point with index k is suited to be the exemplar for data point i; r be the responsibility matrix that reflects the accumulated evidence for how well-suited point k is to serve as the exemplar for point i, taking into account other potential exemplars for point i; and a the availability matrix which reflects the accumulated evidence for how appropriate it would be for point i to choose point k as exemplar, taking into account the support from other points that point k should be an exemplar [32].

These matrices are iteratively updated by the following three equations:

$$r(i,k) \leftarrow s(i,k) - \max_{k' s.t.\ k' \neq k}\{a(i,k') + s(i,k')\} \qquad (3.9)$$

$$a(i,k) \leftarrow \min\{0, r(k,k) + \sum_{i' s.t.\ i' \notin \{i,k\}} \max\{0, r(i',k)\}\} \qquad (3.10)$$

$$a(k,k) \leftarrow \sum_{i' s.t.\ i' \neq k} \max\{0, r(i',k)\} \qquad (3.11)$$

For point i, the value of k that maximizes $a(i,k) + r(i,k)$ either identifies point i as an exemplar if $k = i$, or identifies the data point that is the exemplar for point i. The

[4] Exemplar is similar as centroids in classical clustering, but these centroids are selected from actual data points.

message-passing procedure may be terminated after a fixed number of iterations, after changes in the messages fall below a threshold, or after the local decisions stay constant for some number of iterations.

3.3.3 Clustering Validity Indices

Cluster validity refers to formal procedures to provide an analytical assessment of the amount and type of structure captured by a partitioning, and should therefore be a key tool in the interpretation of clustering results in a quantitative and objective fashion [34, 35].

Two criteria have been proposed for clustering evaluation and selection of an optimal clustering scheme [34]:

1. Compactness: The fitness variance of the patterns in a cluster is an indication of the cluster's cohesion or compactness. Compactness is used as a measure of the variation or scattering of the data within a cluster.
2. Separation: The clusters themselves should be widely spaced. Separation is used to account for inter-cluster structural information.

The basic aim of validation indices has been to find the clustering that minimizes the compactness and maximizes the separation [31].

In general, there are three fundamental criteria to investigate cluster validity: external criteria, internal criteria, and relative criteria [25]. An external assessment of validity compares the recovered structure to an a priori structure. An internal examination of validity tries to determine if the structure is intrinsically appropriate for the data. A relative test compares two structures and measures their relative merit [27].

Although many different cluster validity measures have been proposed [34, 36, 37], the most widely used validity indices are introduced that have direct applicability to

the load profile development. Herein, only hard clustering algorithms are considered avoiding the fact that an object may belong to more than one cluster with different degree of belief (fuzzy clustering), and therefore, fuzzy clustering will be converted to a hard clustering.

3.3.3.1 External criteria

Since external indices are based mainly on prior information of the data, e.g. the optimal numbers of clusters, the indices are used for choosing the best clustering method for a specific data set. This implies that the results of a clustering algorithm are evaluated based on a pre-specified structure, so it will not be considered for the current study.

3.3.3.2 Internal criteria

A good clustering algorithm generates clusters with high intra-cluster homogeneity (compactness), good inter-cluster separation, and high connectedness between neighboring data points [37].

3.3.3.2.1 Cophenetic correlation coefficient

The cophenetic correlation coefficient is used to validate the hierarchy of clustering schemes to measure the degree of similarity between the cophenetic matrix P_c and the proximity matrix P. The cophenetic correlation coefficient index is defined as

$$CPCC = \frac{\frac{1}{M}\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} d_{ij}c_{ij} - \mu_P\mu_C}{\sqrt{\left(\frac{1}{M}\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} d_{ij}^2 - \mu_P^2\right)}} \quad (3.12)$$

where $M = \frac{n(n-1)}{2}$ and μ_P, μ_C are defined as

$$\mu_P = \frac{1}{M}\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} d_{ij}, \quad (3.13)$$

$$\mu_C = \frac{1}{M}\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} c_{ij}, \quad (3.14)$$

where d_{ij} and c_{ij} are the (i, j) elements of matrices P and P_c respectively. The range of CPCC is [−1, 1]; the high value indicates great similarity between P and P_c.

3.3.3.3 Relative criteria

Relative criteria concentrate on the comparison of clustering results generated by different clustering algorithms or the same algorithm but with different input parameters. Herein, the number of clusters k is defined as input parameter. A partition with too many clusters complicates the true clustering structure, therefore making it difficult to interpret and analyze the results. On the other hand, a partition with too few clusters causes the loss of information and misleads the final decision [23].

For a clustering algorithm that requires the input of k from users, a sequence of clustering structures can be obtained by running the algorithm several times from the possible minimum k_{min} to the maximum k_{max}. In this case, to choose the best clustering scheme, the following procedure is performed [26]:

for $k = k_{min}$ to k_{max} **do**

for $i = 1$ to r **do**

Run the clustering algorithm using parameters which are

different from in the previous running;

Compute the value q_i of the validity index;

end for

Choose the best validity index in $\{q_1, \dots, q_r\}$;

end for

3.3.3.3.1 Davies-Bouldin index

The Davies-Bouldin (DB) index is a validity index that does not depend on the number of clusters and the clustering algorithms. To define the DB index, the dispersion measure and the cluster similarity measure need to be defined. The dispersion measure S of a cluster C is defined in such a way that the following properties are satisfied:

1. $S \geq 0$;

2. $S = 0$ if and only if $x = y \ \forall x, y \in C$.

For instance,

$$S_i = \left(\frac{1}{|C_i|}\sum_{x \in C_i} d^p(x, c_i)\right)^{\frac{1}{p}}, \quad p > 0, \qquad (3.15)$$

where $|C_i|$ is the number of data points in cluster C_i, c_i is the center (or representative data point) of cluster C_i, and $d(x, c_i)$ is the distance between x and c_i.

The cluster similarity measure R_{ij} between clusters C_i and C_j is defined based on the dispersion measures of clusters C_i and C_j and satisfies the following conditions:

- $R_{ij} \geq 0$;
- $R_{ij} = R_{ji}$;
- $R_{ij} = 0$ if and only if $S_i = S_j$;
- if $S_j = S_k$ and $D_{ij} < D_{ik}$, then $R_{ij} > R_{ik}$;
- if $S_j > S_k$ and $D_{ij} = D_{ik}$, then $R_{ij} > R_{ik}$;

Here S_i, S_j, S_k are the dispersion measures of clusters C_i, C_j, C_k, respectively, and D_{ij} is the distance (dissimilarity measure) between the two clusters C_i and C_j, which can be defined as the distance between the centroids of the two clusters,

$$D_{ij} = \left(\sum_{l=1}^{d} \left|\boldsymbol{v}_{il} - \boldsymbol{v}_{jl}\right|\right)^{\frac{1}{t}} \qquad (3.16)$$

where $\boldsymbol{v}_i$, $\boldsymbol{v}_j$ are the centroids of clusters C_i and C_j, respectively, and $t > 1$.

A very simple choice for C_i and R_{ij} is

$$R_{ij} = \frac{S_i + S_j}{D_{ij}} \qquad (3.17)$$

Then the DB index is defined as

$$V_{DB} = \frac{1}{k}\sum_{i=1}^{k} R_i, \qquad (3.18)$$

where k is the number of clusters and R_i is defined as

$$R_i = \max_{j \neq i} R_{ij} \qquad (3.19)$$

The smallest $V_{DB}(r)$ indicates a valid optimal partition.

3.3.3.3.2 Dunn's index

The Dunn family of indices was designed to find compact and well-separated (CWS) clusters. The Dunn index is defined as

$$V_D = \min_{1\le i\le k}\left\{\min_{i+1\le j\le k}\left(\frac{D(C_i,C_j)}{\max_{1\le l\le k} diam(C_l)}\right)\right\} \qquad (3.20)$$

where k is the number of clusters, $D(C_i, C_j)$ is the distance between clusters C_i and C_j, and $diam(C_l)$ is the diameter of the cluster C_l. Here, $D(C_i, C_j)$ and $diam(C_l)$ can be defined as

$$D(C_i, C_j) = \min_{x\in C_i, y\in C_j} d(x, y) \qquad (3.21)$$

$$diam(C_l) = \max_{x,y\in C_l} d(x, y) \qquad (3.22)$$

From the definition of the Dunn index, a high value of the index indicates the existence of CWS clusters.

3.3.4 Load Profile Development by Means of Clustering Analysis

Clustering has a long and rich history, so different clustering algorithms have been developed over time. However, none of the clustering is superior to the other, but some are similar to the other. Consequently, since the structure of the data are not known a priori, it is up to the analyst to try competing and diverse approaches to determine a suitable algorithm for the clustering at hand [25]. This task has been undertaken following a thorough review and discussion. Therefore, from the clustering algorithm point of view, k-means is used because of its robustness to provide a specific number of clusters with high similarities between objects in the same group, and low similarities between objects in different groups.

The load profiles developed herein correspond to a set of the total number of meters available during the study period for residential customers. The load profiles will follow similar criteria as in Section 3.2 to obtain the results:

1. Load profile without Day Type distinction but weekday, weekend, or holiday if specified.
2. Load Profile at each Day Type of the week: Holidays, Mondays, Tuesdays, Wednesdays, Thursdays, Fridays, Saturdays, and Sundays.

In addition, the load profiles based on clustering can be obtained at each stratification level, if specified.

Similarly as before, as an example of the construction of load profiles based on the customers' behavior similarities, load profiles for summer corresponding to the months of June, July, August, and September of 2012 are calculated. The smart meters belong to the group of Min15Ch2. There are 541 smart meters in total, and they were clustered considering K = 5 to have resemblance to the load profiles developed in Section 3.2. The following was the result: 191 meters belong to Group 1, 164 meters to Group 2, 114 meters to Group 3, 57 meters to Group 4, and 15 meters to Group 5. Figure 3-20 to Figure 3-29 show the Load Profiles for Weekdays and Weekends during Summer 2012 for all 5 groups.

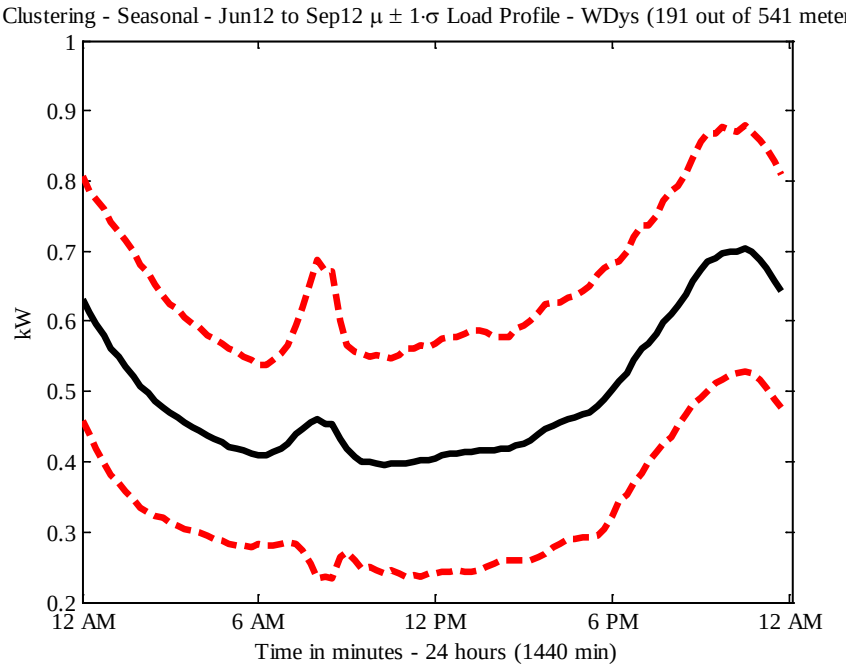

Figure 3-20 Weekdays Load Profile – Mean ± Std. Dev. for Group 1

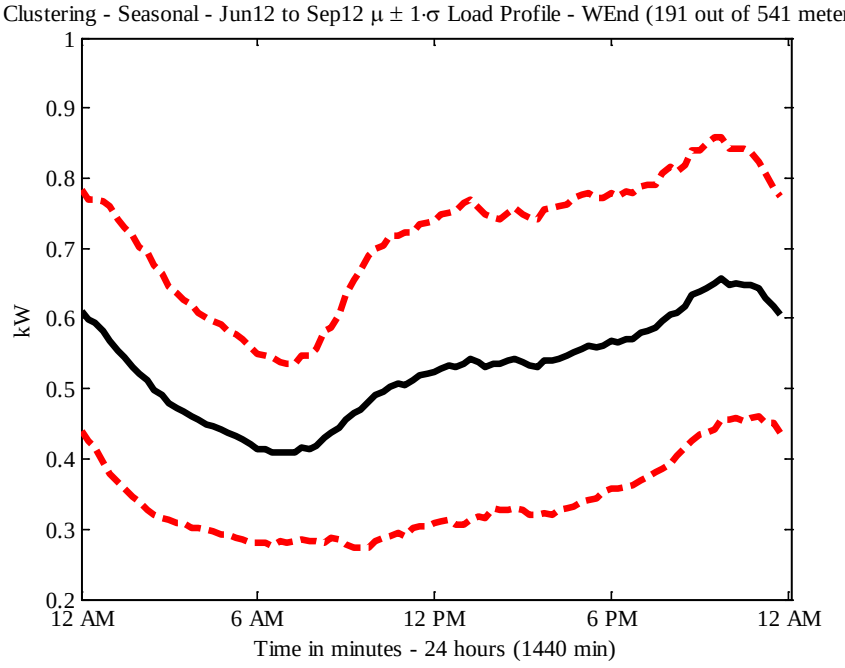

Figure 3-21 Weekends Load Profile – Mean ± Std. Dev. for Group 1

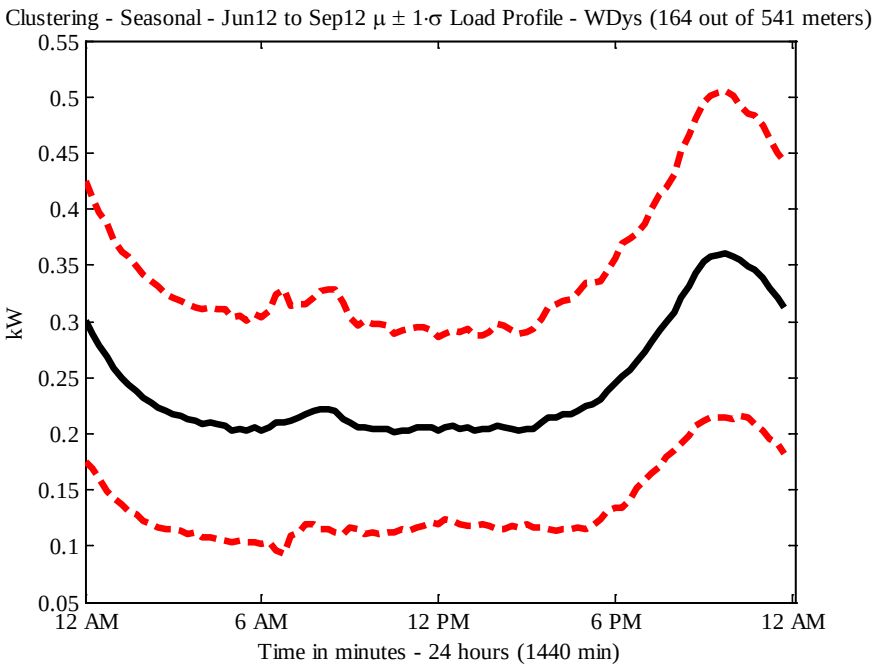

Figure 3-22 Weekdays Load Profile – Mean ± Std. Dev. for Group 2

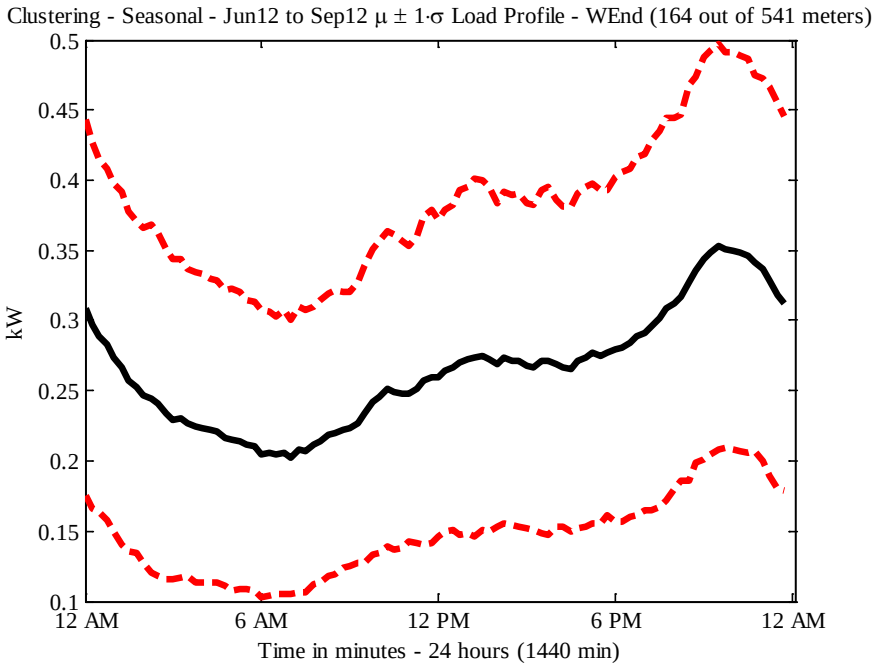

Figure 3-23 Weekends Load Profile – Mean ± Std. Dev. for Group 2

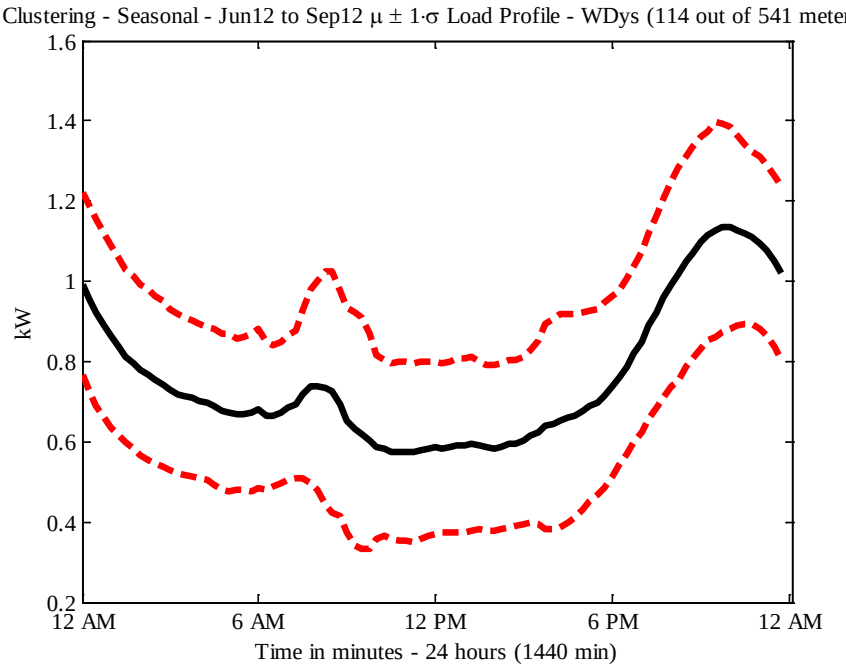

Figure 3-24 Weekdays Load Profile – Mean ± Std. Dev. for Group 3

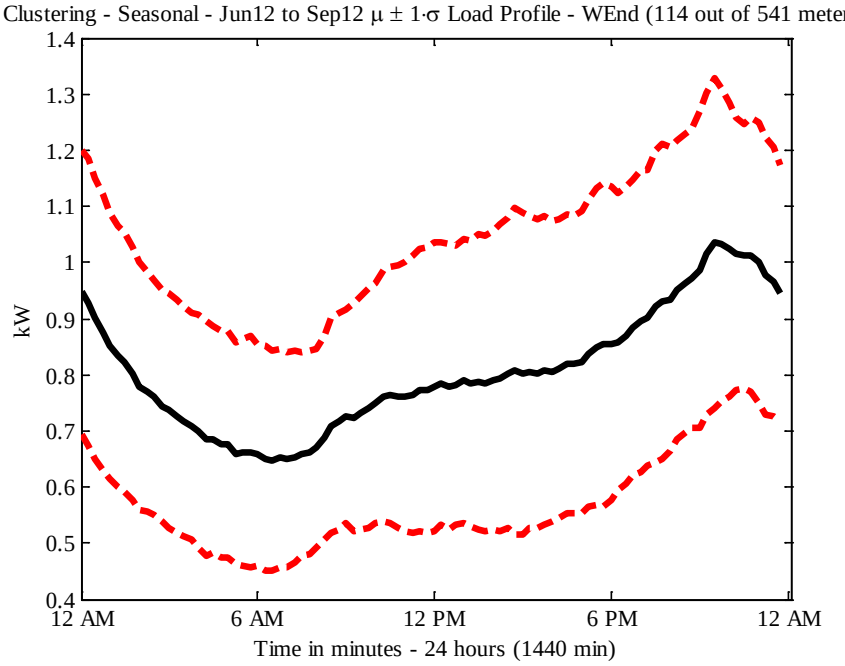

Figure 3-25 Weekends Load Profile – Mean ± Std. Dev. for Group 3

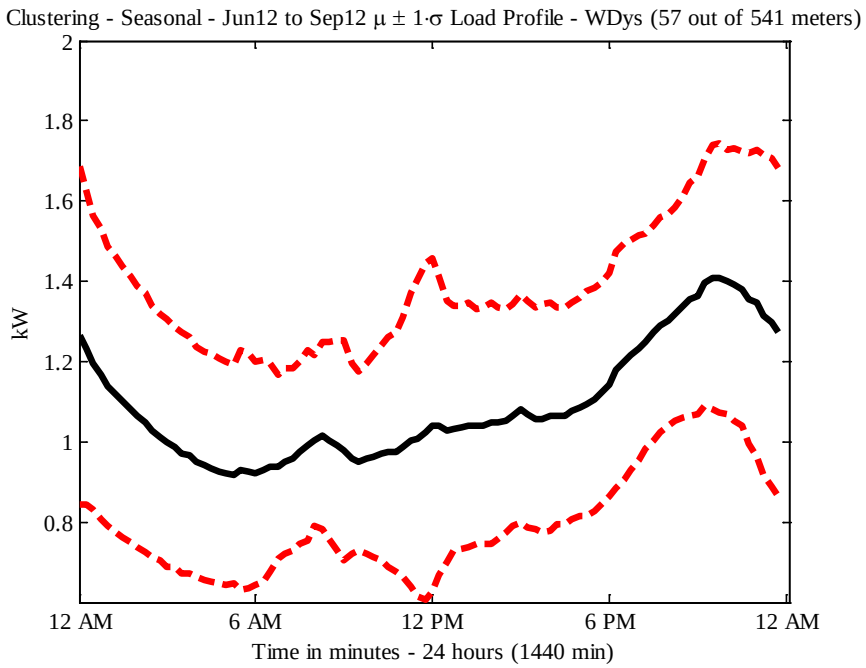

Figure 3-26 Weekdays Load Profile – Mean ± Std. Dev. for Group 4

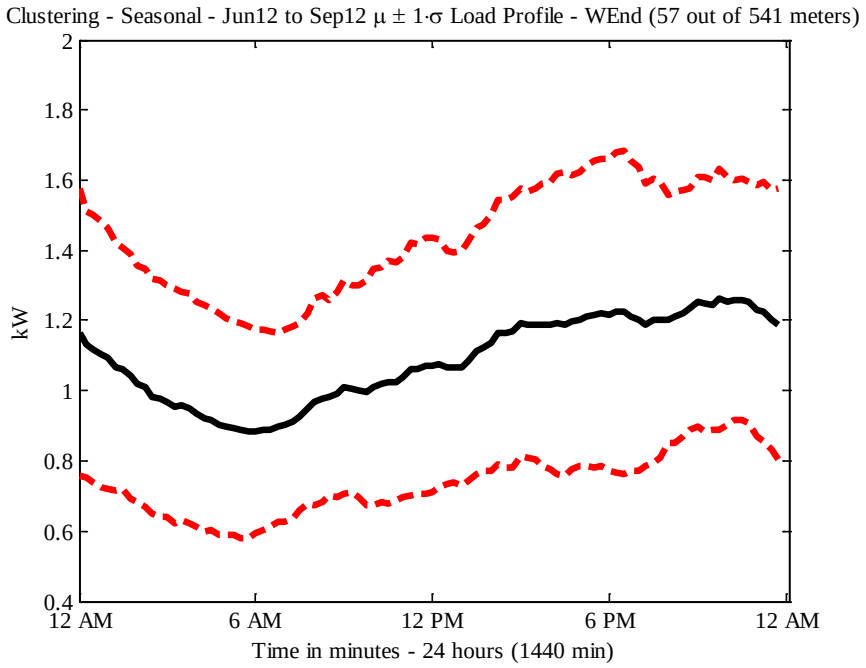

Figure 3-27 Weekends Load Profile – Mean ± Std. Dev. for Group 4

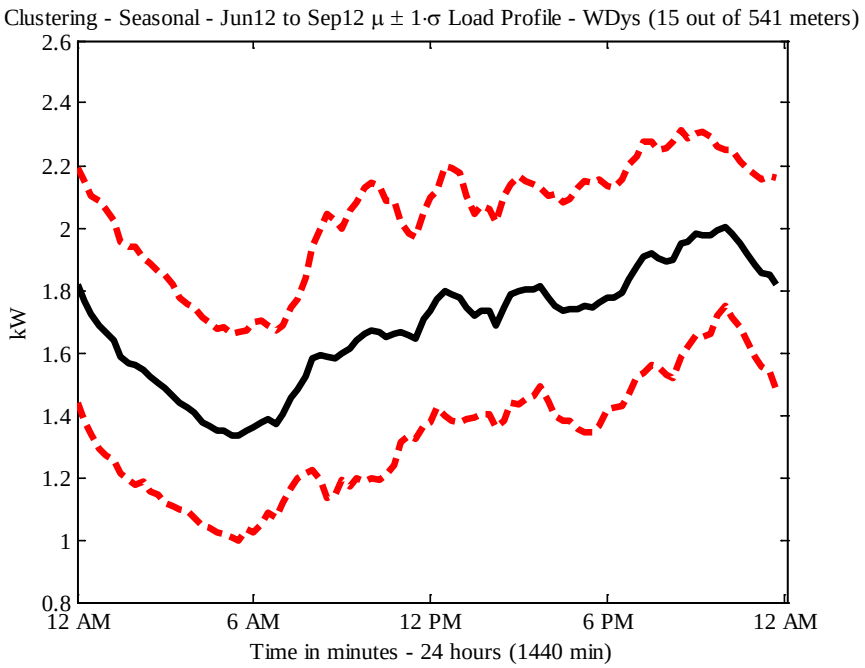

Figure 3-28 Weekdays Load Profile – Mean ± Std. Dev. for Group 5

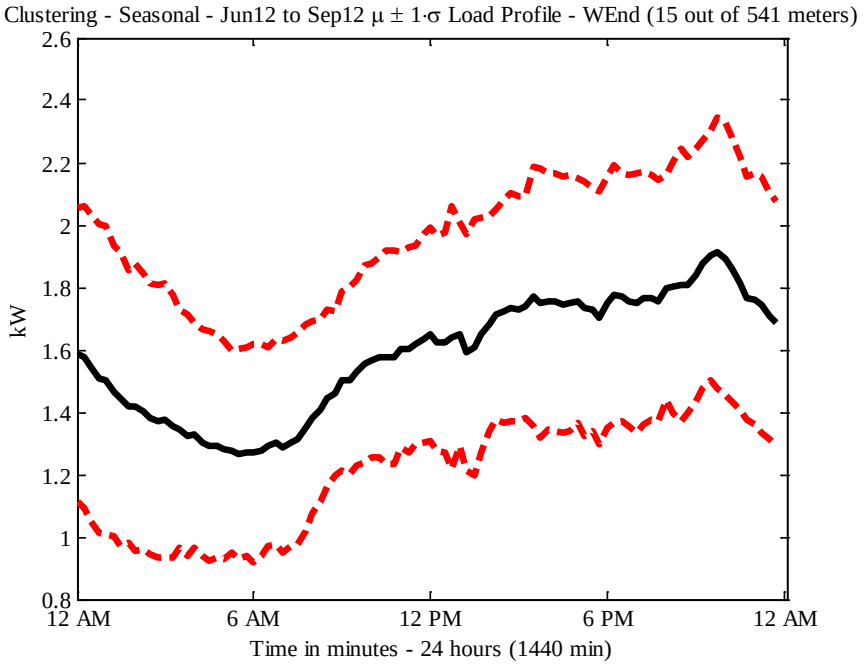

Figure 3-29 Weekends Load Profile – Mean ± Std. Dev. for Group 5

To provide examples of Load Profiles for Day Type considering customers' behavior similarities, only Group 1 is plotted for the sake of clarity from Figure 3-30 to Figure 3-37. Nevertheless, load profiles can be calculated for all groups.

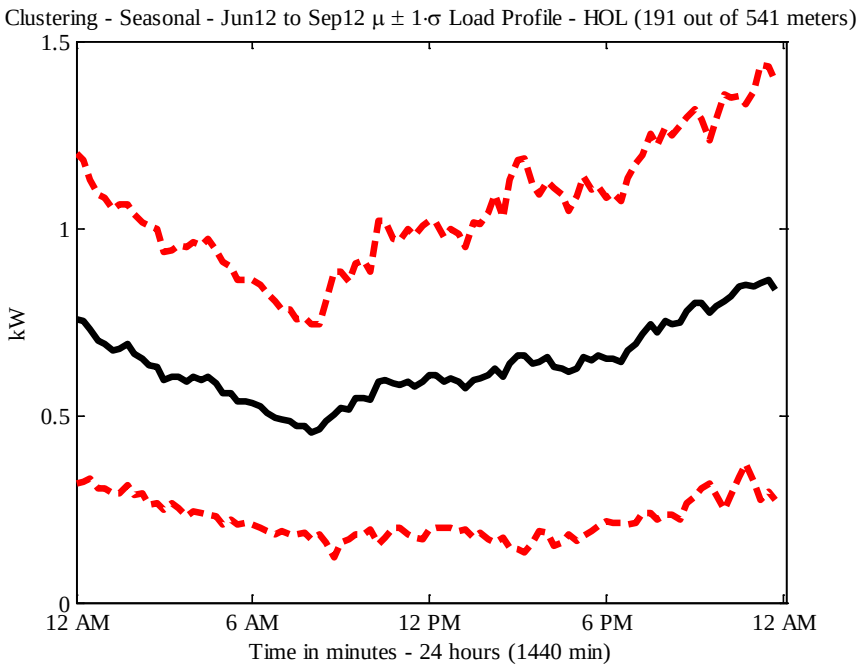

Figure 3-30 Holidays Load Profile – Mean ± Std. Dev. for Group 1

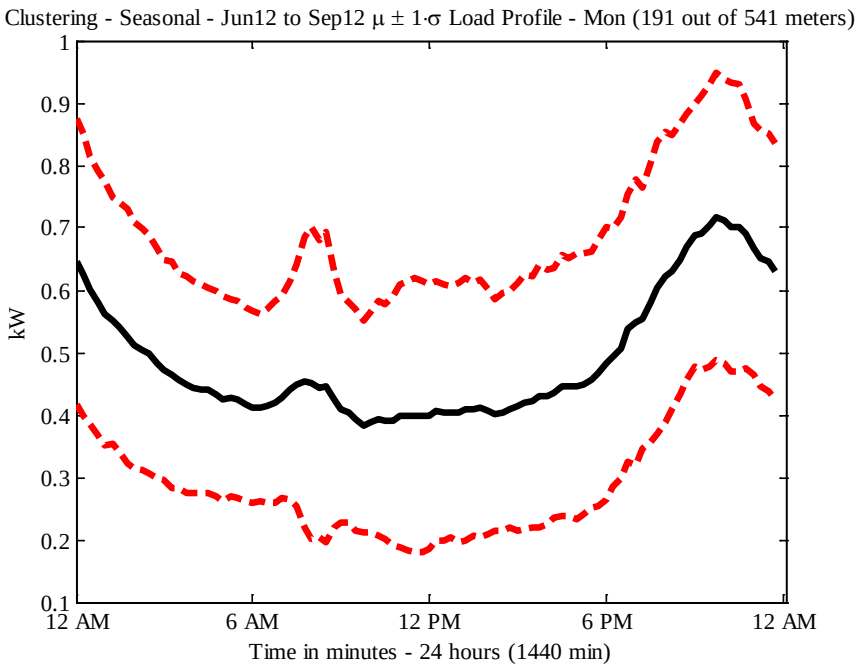

Figure 3-31 Mondays Load Profile – Mean ± Std. Dev. for Group 1

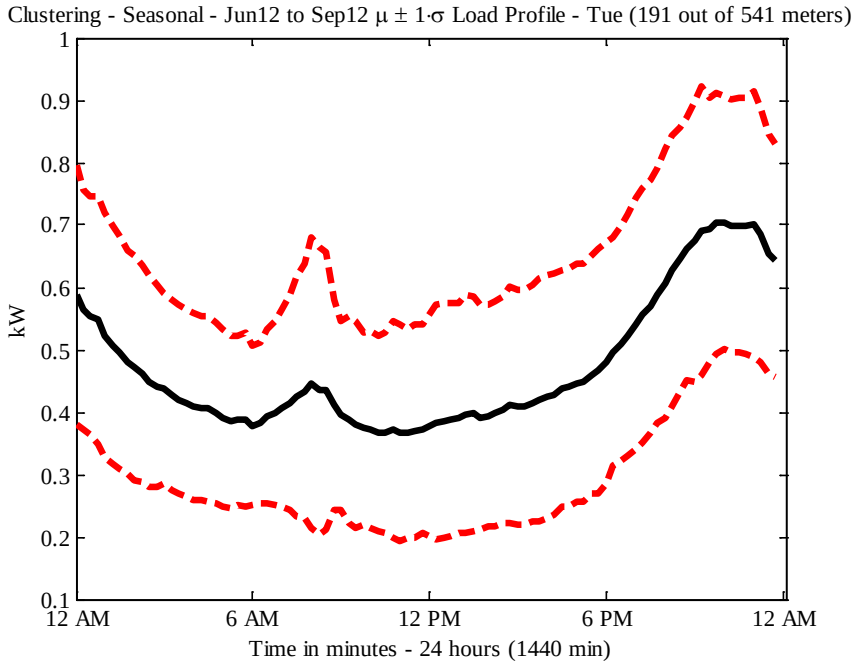

Figure 3-32 Tuesdays Load Profile – Mean ± Std. Dev. for Group 1

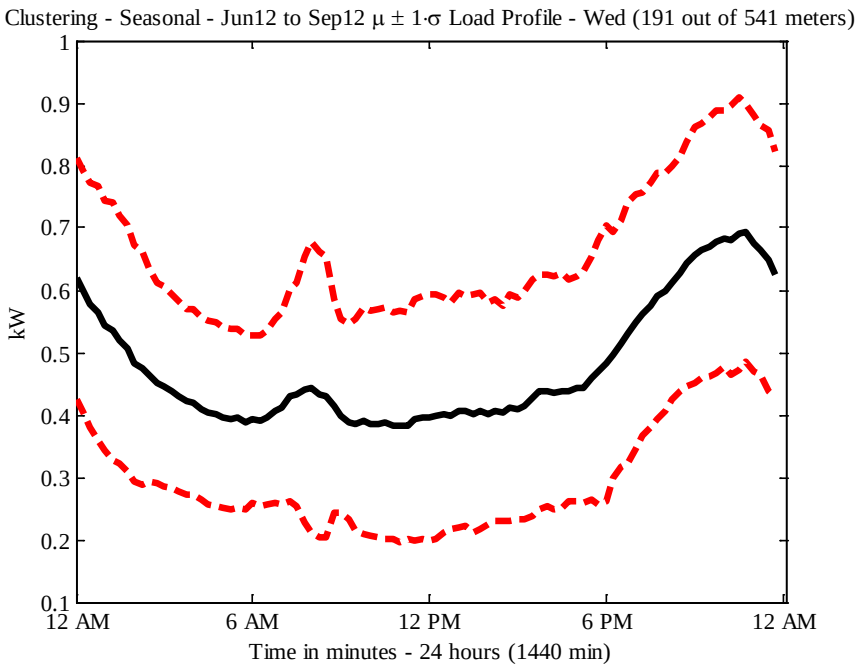

Figure 3-33 Wednesdays Load Profile – Mean ± Std. Dev. for Group 1

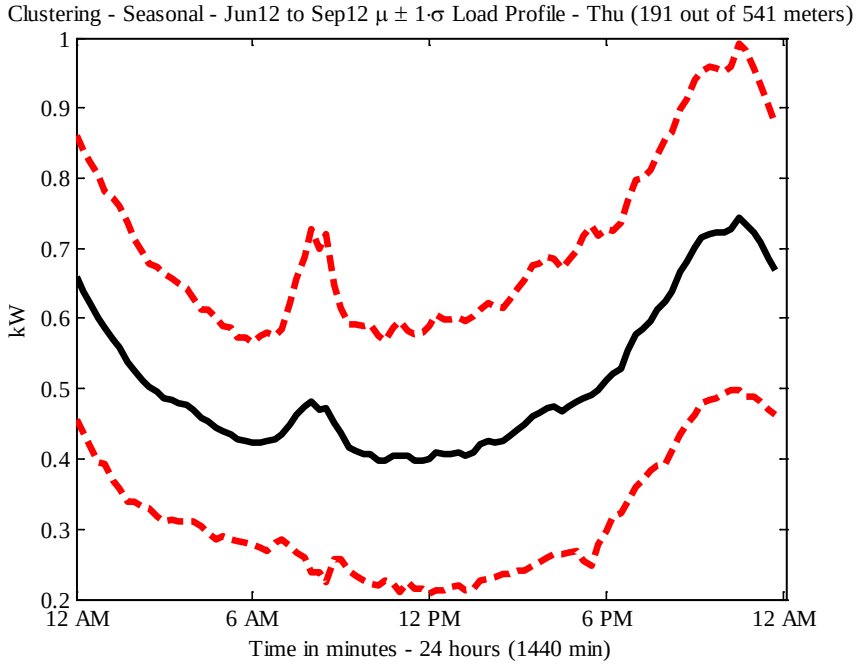

Figure 3-34 Thursdays Load Profile – Mean ± Std. Dev. for Group 1

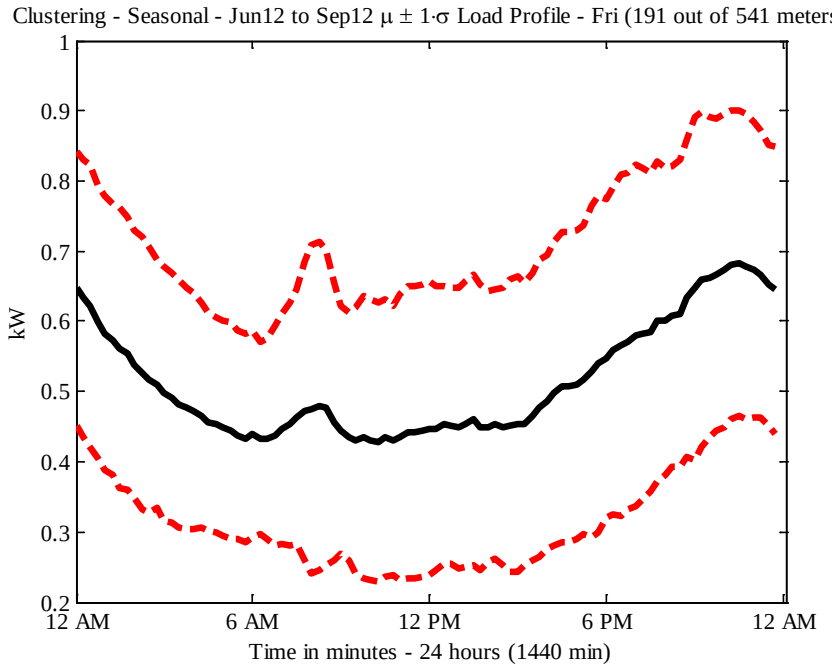

Figure 3-35 Fridays Load Profile – Mean ± Std. Dev. for Group 1

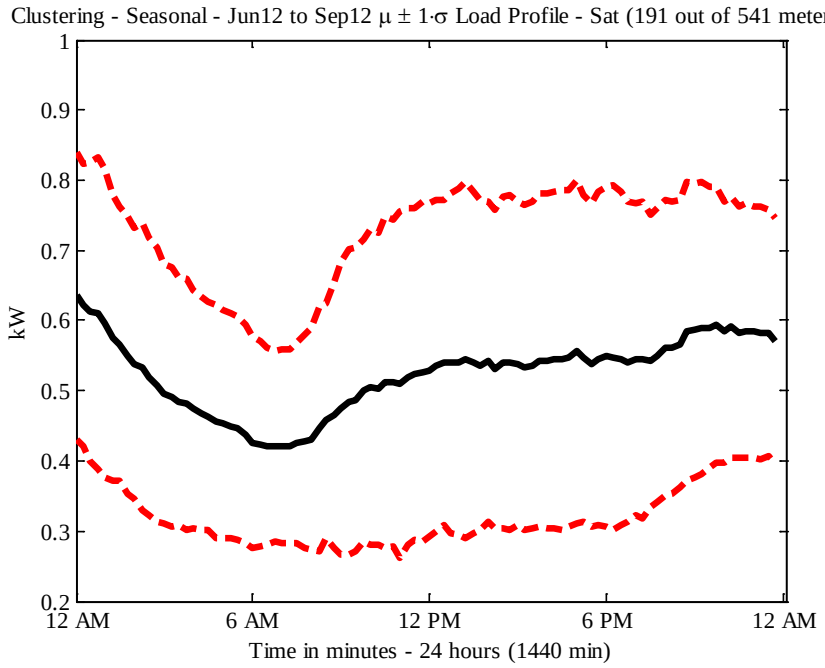

Figure 3-36 Saturdays Load Profile – Mean ± Std. Dev. for Group 1

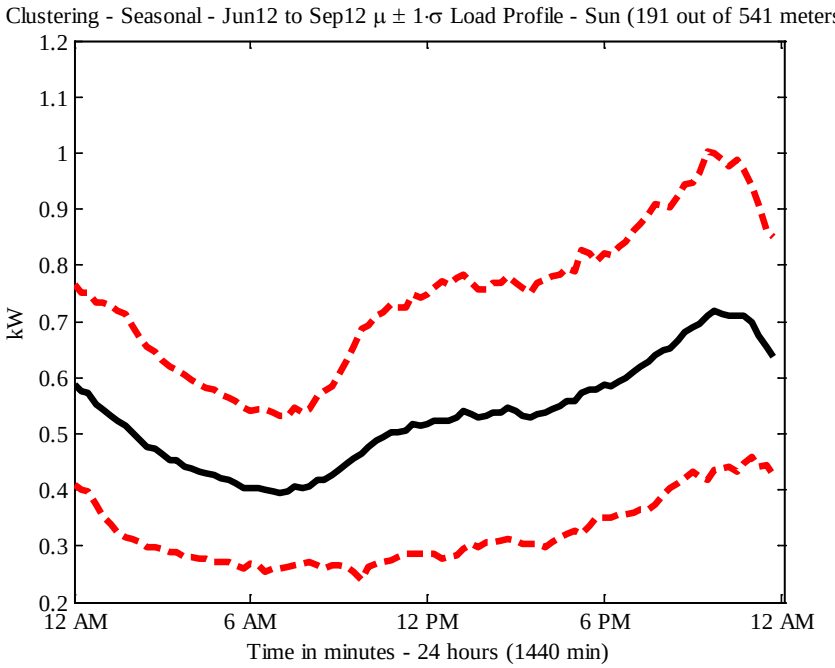

Figure 3-37 Sundays Load Profile – Mean ± Std. Dev. for Group 1

Chapter 4

AMI Data to Enhance the Performance of Load Forecasting

Load forecasting is an essential task for multiple utility business processes including power generation, power trading, capacity planning, and demand management. For instance, demand forecasting is useful for planning and purchasing power supply by utilities, scheduling equipment maintenance, and providing an early warning of potential load curtailment or advance pricing information. Since load forecasting accuracy has a significant impact on scheduling, operation, and control of the utility grid, it is crucial to accurately know the total and local system demand for the following minutes, hours, and days. This is the domain of Very Short Term Load Forecasting (VSTLF), and Short Term Load Forecasting (STLF). When the load forecasting is concerned with the prediction of longer time horizons, they are usually categorized as Mid Term Load Forecasting (MTLF), and Long Term Load Forecasting (LTLF). Table 4-1 shows examples of this LF classification.

Table 4-1 Load Forecasting Classification

Load Forecasting	Load Data Resolution[5]	Horizon	References
VSTLF	1 minute	30 minutes	[39]
	1 minute	10 to 30 minutes	[40]
	5 minutes	1 hour	[41]
STLF	1 hour	24 hours – 3 weeks	[42]
	½ hour	1 day	[43]
MTLF	½ hour	31 days	[44]
LTLF	½ hour	10 years	[45]

Though technical literature presented a wide range of methodologies and models to improve the accuracy of load forecasting, most of them are based upon aggregated power consumption data at the system (corporate) level with little or even no information

[5] Load data resolution at the net system level.

regarding power consumption profiles of different classes of customers. With the deployment of Advanced Metering Infrastructure (AMI), an avalanche of new energy-use information becomes available. Unlike traditional aggregated system level load forecasting, the AMI data introduces a fresh perspective to the way load forecasting can be performed, ranging from very-short term load forecasting (STLF) to long-term load forecasting (LTLF) at the system level, regional level, feeder level, or even down at the consumer level. One of the most critical steps to realizing these benefits is to develop data management and analysis processes to transform smart meter data into useful information for load forecasting and other applications.

This dissertation addresses the efforts involved in using AMI data from residential customers as an example to the Load Forecasting problem, focusing particularly on utilizing sub-hourly interval data with a time horizon up to one day ahead and assessing the possibility of applying unsupervised learning to identify customers' consumption patterns first, and then developing load forecasting models at each identified group. Load forecasting can take full advantage of clustering methods because grouping load profiles based on consumption behavior similarities will reduce the variability of predicted load, and therefore, the forecasting error.

4.1 A Review on Load Forecasting Techniques

Since the main interest of this work is to utilize AMI data for Load Forecasting, this dissertation focuses on Load Forecasting of sub-hourly data with a time horizon up to one day ahead. Out of the four categories, VSTLF and transition to STLF could serve well for this purpose.

Many different techniques have been introduced for STLF with system level data. A comprehensive review on this subject can be found in [46-49]. In general, time series models, traditional econometric models, artificial neural network-based models, fuzzy

logic-based models, nonparametric/semi-parametric regression-based models, hybrid models (i.e. combination of different models like neuro-fuzzy models, among others), and judgmental forecasting models have been applied with relatively high accuracy to STLF.

On the other hand, only a few articles have been published on VSTLF [40] using system level data. The following gives a general overview of VSTLF approaches.

In [39], the authors performed a forecast of the next 30 minutes of real time (moving window) load on intervals of 1 minute. They utilized two-stage fuzzy logic (FL) for training and on-line forecasting. In training, patterns are recognized and stored in a database of one-minute filtered historical load data. In on-line mode, input load data is compared with the patterns to predict the load. Also, a fully-connected feed-forward neural network with 38 inputs (past 30 load values, 4 time components, 4 load parameters), 16 outputs (the Karhunen–Loève transformed values of the next 30 load values), and two hidden layers (30 and 25 units, respectively) with sigmoidal activation functions was used for predicting the load. Finally, an auto-regressive (AR) model of the next minute load but whenever a new prediction is generated, it is treated as the new actual load datum until 30 data are generated. Authors concluded that FL and NN performance are much superior to AR-based forecaster.

In [41], the authors used separate neural networks applied to wavelet decomposed filtered load data. NN results are combined to produce the final forecast. The forecasting output was forecast of 1 hour in 5-minute steps in a moving window manner.

In [50], the authors used “parsimoniously designed” neural networks to forecast relative increments in load based on the recent load pattern. Each forecaster was trained using the data from day 1, and then used to forecast the 20-min ahead load of day 2.

They predicted 8 values of load for the time leads from 20-90 minutes in 10-minute increments.

For VSTLF to STLF approach, in [40], the authors presented and compared several univariate methods including a comparison with an additional approach based on weather forecasts. Best results were achieved by a double seasonal adaptation of Holt-Winters' exponential smoothing. In addition, beyond VSTLF, combining methods based on weather forecast with the Holt-Winters' adaptation are promising to forecast load beyond an hour ahead. The forecast horizon was up to 30 minutes ahead in 1-minute step.

In contrast, load forecasting utilizing real-world smart meter data can be summarized as follows:

In [51], the authors proposed a "short-term multiple load forecasting (STMLF)" model which combines individual load time-series into a succinct model for forecasting many loads with a single model and the use of anthropologic and structural data within STMLF to tackle the problem of the high volatility in dynamics for an individual customer that makes forecasting for each individual load difficult. While multiple linear regression and artificial neural networks were the main forecasting engines, the authors suggested that ANN and SVM will be suitable for STMLF.

In [38], the authors discussed time series approaches that can be used to characterize individual customer demand load profiles, even though time-series have rarely been used at an individual dwelling level. Fourier transforms and Gaussian processes were devised suitable techniques to accomplish this task when applied to half-hourly electricity demand on a daily basis for each individual customer.

In [52], the authors examined six methods, successfully used for forecasting energy demand on a large scale, to forecast the load on a smaller scale similar to the

load of a single transformer. Artificial neural network, auto-regressive, auto-regressive moving average, auto-regressive integrated moving average, fuzzy logic, and wavelet neural networks were utilized for day ahead, and week ahead electric load forecasting in two different scenarios, with 90 houses, and with 230 houses. The authors concluded that at a small scale, the noise and chaotic behavior have a great impact on the forecast accuracy.

In [53], the author proposed forecasting functional time series applied to intra-day household-level load curves using smart meter data via two methods: Functional Wavelet-Kernel (FWK) and Clustering-Based FWK. Both approaches are identical, except the latter identifies a common pattern between days at each individual customer following the idea of the similar-day approaches but through an unsupervised classification method, and then it utilizes FWK to perform one-day ahead forecasting of each individual customer. It is also stated that household loads are very volatile, which makes household-level forecasting difficult to solve.

Taking into consideration the above description, it is clear that despite the fact that load forecasting is a challenging task by itself at any level and at any time horizon, it is more difficult to forecast load at the household level with fine granularity data. Forecasting individual load consumption for each customer at an electric utility company will require large computing resources due to the volume of data. Moreover, to apply any load forecasting technique for each household will require identifying what drives individual load consumption behavior in great detail to be able to capture adequately its complex dynamics.

It also reveals two general trends of using smart meter data in Load Forecasting: (1) Forecast individual household loads, and (2) aggregate all the loads and construct a single forecasting model for the system load. This dissertation proposes a different

approach that takes into account smart meter data (lower level), to forecast load at a larger level, e.g. system level, that considers individual volatile loads grouped based on consumption behavior similarities, and then develop load forecasting models at each identified group. Since the interest is to utilize AMI data for Load Forecasting, this dissertation is focused on Load Forecasting of sub-hourly data with different time horizons up to one day ahead.

Out of the several load forecasting techniques, neural network is selected to illustrate the proposed approach.

4.2 Neural Network-Based Load Forecasting

Neural networks (NNs) are well accepted in practice and used by many utilities [54]. Moreover, since NNs can approximate any continuous function, they can be seen as a multivariate, nonlinear, method that can model complex nonlinear relationships. In addition, NNs are data-driven methods, and therefore well suited for using them with smart meter data.

When designing a neural network-based forecasting model, the first step is selecting an appropriate architecture. Although there are many types of NNs [55], in reality, most NN-based load forecasting models utilize a feed-forward multilayer perceptron (MLP) with satisfactory results in terms of accuracy [48]. In a typical MLP, neurons are organized in layers: one input layer, one or more hidden layers, and one output layer. Once the MLP-NN architecture is selected, one must decide the number of input nodes, number of hidden layers and the type of activation function, and the number of output nodes.

Considering that the purpose is to forecast the load with a horizon up to one day ahead at a resolution equal to the meters' resolution, there are essentially two ways on doing so:

1. Use the NN model to forecast one step ahead. Note that for leading times larger than the time interval considered, the one-step ahead forecasts can be iteratively used as inputs in order to generate multi-step predictions [40, 43].
2. Use the NN model to forecast multi-steps ahead. By using a system of NNs, one for each time interval or a large NN with many outputs corresponding to each interval for a full day ahead forecast [48].

In this study, method 1 is explored. Therefore, the neural network-based forecasting model's architecture is an MLP with one hidden layer with hyperbolic tangent activation function, and one linear output neuron. The hyperbolic tangent function *g* is given by [56]:

$$g(h) = \frac{2}{1+e^{-2h}} - 1 \qquad (4.1)$$

The parameters of this NN are the weights associated with the connections from the input nodes to the hidden layer, and the weights for the connections from the hidden layer to the output node. The estimation of the network parameters is called "training the NN", and its purpose is to find the weighting matrices that minimize a loss function [48].

The Levenberg-Marquardt approach is used to train the model. This approach is suitable for training medium-size NNs with low mean square error. One of the key problems in NN application is to select the number of hidden neurons in the hidden layer which affects the learning process and forecasting capability of the network. An approach similar as in [57] is adopted to overcome this problem. The method starts by choosing a small number of hidden neurons and gradually increases this number. Each time, the model is trained, and a forecast error from the testing set is recorded for comparison. The process stops at an optimal number of hidden neurons when the error decreases to an acceptable threshold or no significant improvement is observed as the number of hidden neuron increases. Another issue that occurs during neural network training is called

overfitting where the NN loses its generalization ability. This problem is tackled by using regularization [58]. Without getting into much detail, it was determined that using 20 hidden neurons and a regularization parameter of 0.9 performed well in the study. Moreover, because the estimation of the weighting matrices, from input to hidden layer and from hidden layer to output, is sensitive to the choice of initial values, each model was estimated 20 times from random initial values. Then, the best model was determined by calculating the out-of-sample MAPE. The lower the out-of-sample MAPE is, the better the model.

Last but not least is the selection of input variables. A highly significant model term does not necessarily translate into good forecasts [59]. To select appropriate input variables, it is primarily important to understand which factors contribute to the load consumption. It is well understood that load consumption is mainly driven by temperature, with seasonal patterns. Once all the possible input variables are identified, one can begin with the full model including all the variables. Then, the predictive capacity of each variable is tested independently by dropping each term from the model while retaining all other terms. Omitted variables that led to a decrease in MAPE were left out of the model for subsequent test [59]. Through the above mentioned process, the variables which are used in the model are:

Smart Meter Data Variables

- Interval (sub-hourly) Load Readings
- Lagged Load Readings at a sub-hourly resolution: last 3 hours same day, last 3 hours day before (plus same hour day before), last 3 hours previous week (plus same hour previous week)

Calendar Variables

- Day of the Week

- Holiday
- Month of the Year

Weather Variables

- Temperature Variables: Temperature was interpolated between neighboring values to obtain measurements at a resolution similar as the smart meter data (e.g. 15 or 30 minutes). Only historical values were considered and no temperature forecasting was used as input for the load forecasting.

4.3 Forecasting Application

4.3.1 Data Collection

Real-world smart meter data for residential customers of two different electric utility companies (1) from the United States, and (2) from Ireland are used in this study.

Starting in 2009, the Consolidated Edison Company of New York, Inc. (Con Edison) initiated a Smart Grid Project aimed at deploying a wide-range of grid-related technologies, including automation, monitoring and two-way communications, to make the electric grid function more efficiently and enable the integration of renewable resources and energy efficient technologies [60]. Twenty one months of 15-minute load data from February 2012 to October 2013 have been used to validate the proposed approach for STLF. The first 12 months of the data were used for training the model parameters, and observations from the last 9 months were used for model evaluation.

The Commission for Energy Regulation (CER) made publicly available full data sets in an anonymized format of the recent “Electricity Smart Metering Customer Behaviour Trials” in Ireland recorded from July 14, 2009 to December 31, 2010 with over 5,000 Irish homes and businesses [61]. The data were obtained via the Irish Social Science Data Archive (ISSDA) [62]. The smart meters collected the electricity consumption at a resolution of 30 minutes. Although there are three available service

classes from the data set: (1) Residential, (2) Small-to-Medium Enterprises (SMEs or Commercial), and (3) Other, only Residential customers were considered for this study, and without distinction of any tariff program. Seventeen months of half-hourly load data from August 2009 to December 2010 have been used to validate the proposed approach for STLF. Similarly, the first 12 months of the data were used for training the model parameters, and observations from the last 5 months were used for model evaluation.

Data preprocessing was performed to verify the quality of the smart meter data following the concepts introduced in Chapter 2. Records with missing or incomplete data can have significant impact on the accuracy of the predictive model. Noisy/incomplete AMI data reads will lead to an inaccurate sequence of load forecasts. These records must be completed, corrected, or eliminated so that the ultimate predictive model is as accurate as possible.

In addition to these data sets, temperature data at both locations for their respective time periods were obtained from wunderground.com.

With all these data, point forecasts were generated using a rolling forecast for horizons varying from 15 minutes or 30 minutes up to one day ahead.

4.3.2 Smart Meter Load Data Grouping Based on Clustering

Figure 4-1 and Figure 4-2 shows a daily “system” load profile composed by the aggregation of all residential customers, and also a single residential customer load profile on July 17, 2012, and December 25, 2009, for the two datasets respectively. It can be observed that the consumption across the day is very different from that of an aggregated system load profile. Each household will have an individual daily load curve, though each will be different because each home has different appliances and is occupied by people with different schedules and usage preferences.

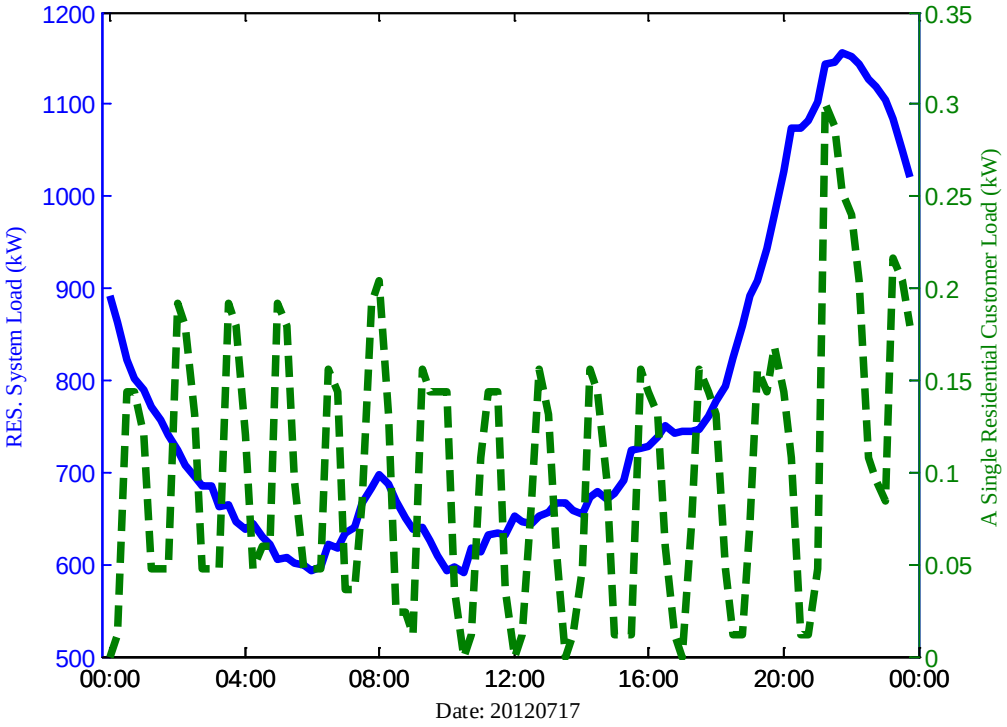

Figure 4-1 Daily load profile for Residential Customers for Residential System Demand, and a Single Residential Customer Across a 24-hour Period on July 17, 2012

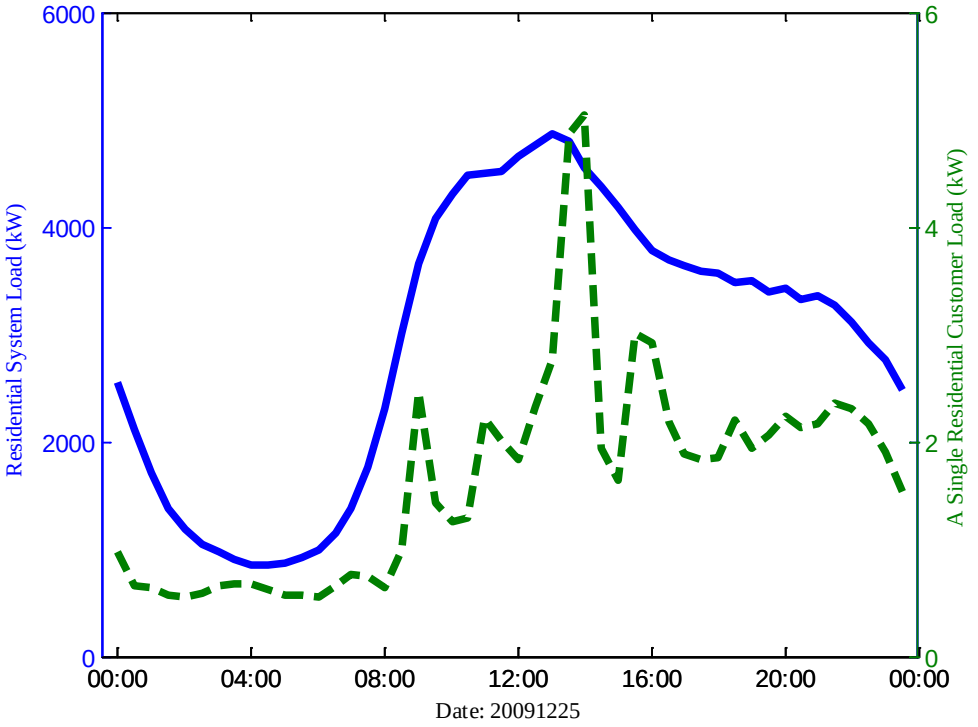

Figure 4-2 Daily load profile for Residential Customers for Residential System Demand, and a Single Residential Customer Across a 24-hour Period on December 25, 2009

Figure 4-3 and Figure 4-4 show six different residential customer profiles (at random) on April 28, 2012, and August 12, 2009 for both datasets respectively. It can be seen that the load pattern consumption changes across every customer. Furthermore, load consumption can change on a daily basis even for a single customer as shown in Figure 4-5 and Figure 4-6.

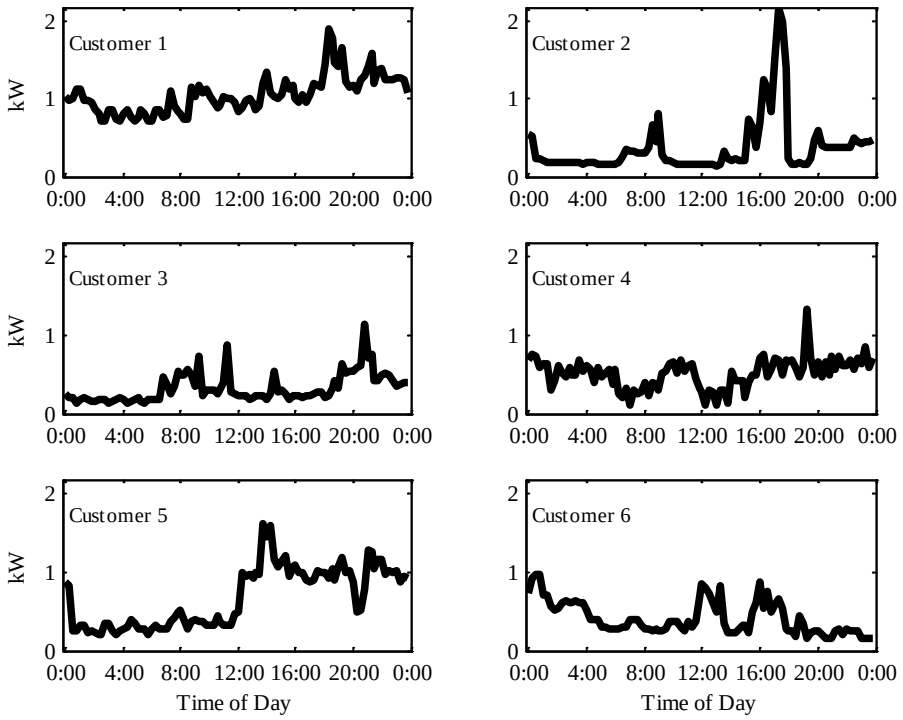

Figure 4-3 Daily load profiles for Six Residential Customers Chosen at Random

Illustrating Variation Between Household Consumers on April 28, 2012

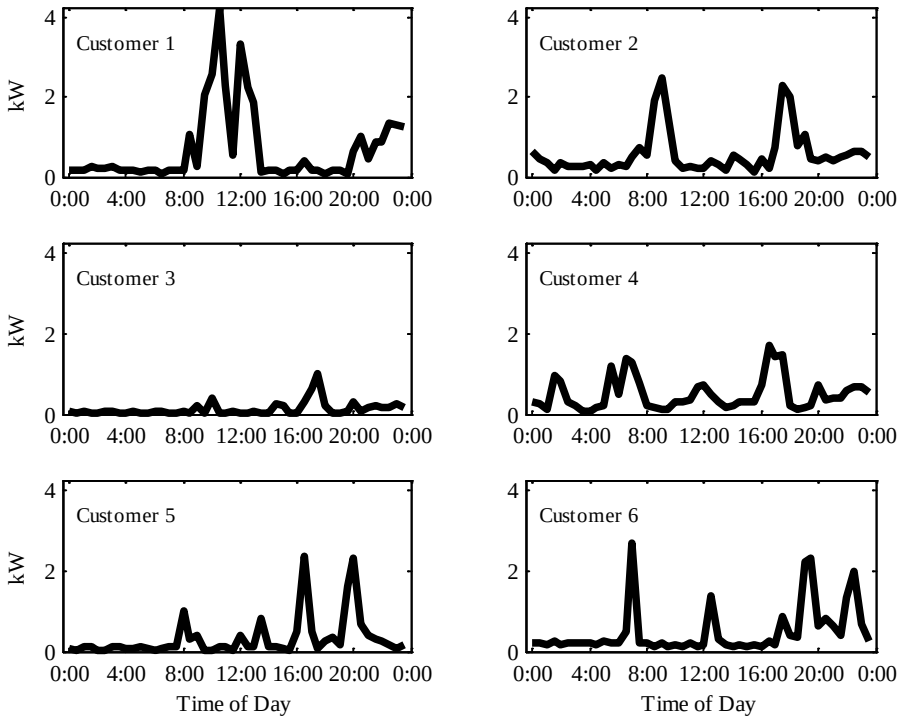

Figure 4-4 Daily load profiles for Six Residential Customers chosen at Random Illustrating Variation Between Household Consumers on August 12, 2009

It should be clear by now that load consumption differs in both, magnitude and time of use, depending on lifestyle, weather, and many other factors, so what it can be achieved by means of clustering is to group load customers in a meaningful way, taking into consideration these inherent daily and intra-daily variations [38] that can/will improve current practice on load forecasting.

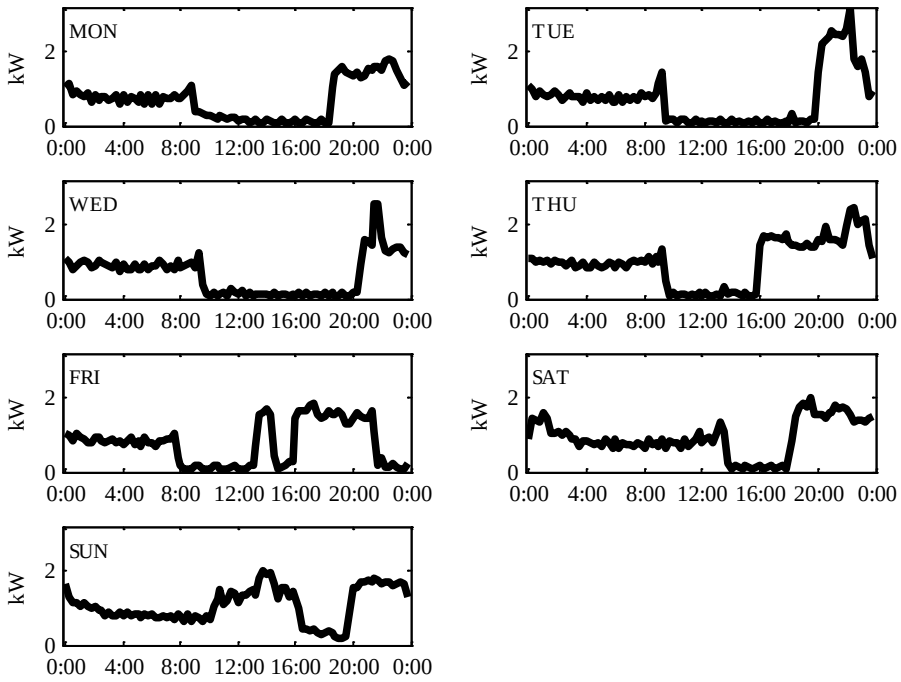

Figure 4-5 Daily Load Profiles for a Single Customer Chosen at Random Over a Weekly Period

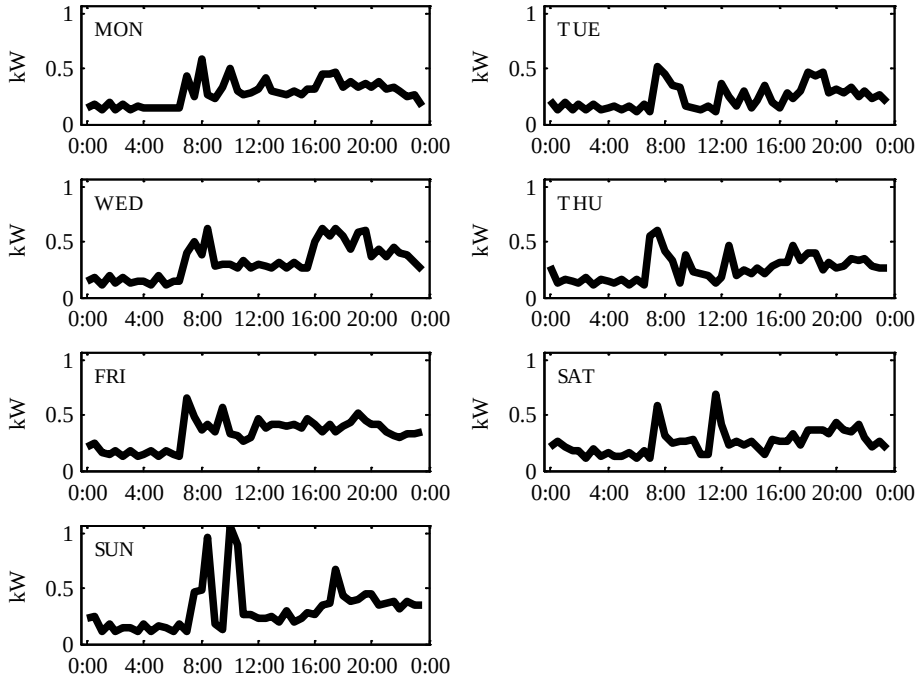

Figure 4-6 Daily Load Profiles for a Single Customer Chosen at Random Over a Weekly Period for Dataset 2

4.3.3 Clustering Implementation for Load Pattern Grouping

The following is a description of the methodology implemented in this study that works well for load pattern grouping to be used in a load forecasting application:

Start by selecting the season of the year where the load peak occurs during the whole year, and do the following for each of the *m* customers:

- Divide each day into 5 segments corresponding to main intraday consumption behavior patterns.
- Obtain an average consumption at each day of the week. It will represent the average consumption pattern of a single customer every day of a typical week.

- Normalize the load in a range of 0 to 1 to emphasize grouping the customers according to who contributes to the total consumption at a certain time of the day, a.k.a. coincident demand. Another benefit of doing so is to obtain a more equally distributed number of customers at each cluster.

The data points can be arranged in the *m-by-n* data matrix *D*, where *m* is the number of meters, and *n* is the number of features. Therefore, the dimension of *D* is equal to 35 (5 segments per day for the 7 days of a typical week). At this moment, k-means is applied with *k* ranging from 1 to 12, and with 1,000 repetitions to overcome the curse of local minimum.

Now, rather than using any clustering validity index to decide on a suitable number of clusters, a different venue is pursued. Since the ultimate goal is to improve load forecasting based on grouping customers based on their consumption behavior, the number of clusters is evaluated based on a metric of how well it performs when forecasting the load. In this dissertation, the Mean Absolute Percentage Error (MAPE) is utilized to measure the forecasting performance, and therefore the MAPE will be used to determine how many clusters are adequate.

4.3.4 Forecasting Results

In this research, the application of clustering to determine groups of customers considering load consumption similarities is studied as an aid to improve the performance of load forecasting at the system level, but with load data at the household level from smart meters. To confirm the findings from this approach, the study was performed in two completely different datasets (1) from a utility in USA, and (2) from a utility in Ireland. As explained in the preceding sections, 78 independent NN-based forecasting models were constructed at each of the groups when *k* was varied from 1 to 12 utilizing the two

datasets. To determine the “optimal” number of clusters, the monthly out-of-sample MAPE was evaluated. Then, the mean MAPE was calculated for the whole testing period.

For dataset 1, it was determined that 3 clusters give a reduction in MAPE of approximately 0.5 % with respect to its counterpart with 1 cluster only, at one-day ahead forecasting. Figure 4-7 depicts this result on predicting the load at different lead times: 15min ahead, 30min ahead, 1h ahead, 2h ahead, …, 24h ahead for 6 clusters only, although it was calculated for 12 clusters, only 6 are depicted for the sake of clarity. Figure 4-8 shows the average load profiles for each one of the three clusters during July 15, 2013 to July 21, 2013. It can be noticed that the load profiles are quite different among different clusters.

For dataset 2, it was determined that 4 clusters give a reduction in MAPE of approximately 1.07 % with respect to its counterpart with 1 cluster only, at one-day ahead forecasting. Figure 4-9 depicts this result for 6 clusters only, for the sake of clarity, on predicting the load at different lead times: 30min ahead, 1h ahead, 2h ahead, …, 24h ahead. Figure 4-10 shows the average load profiles for each one of the four clusters during December 20, 2010 to December 26, 2010. It is obvious that the load profiles are quite different among different clusters as well.

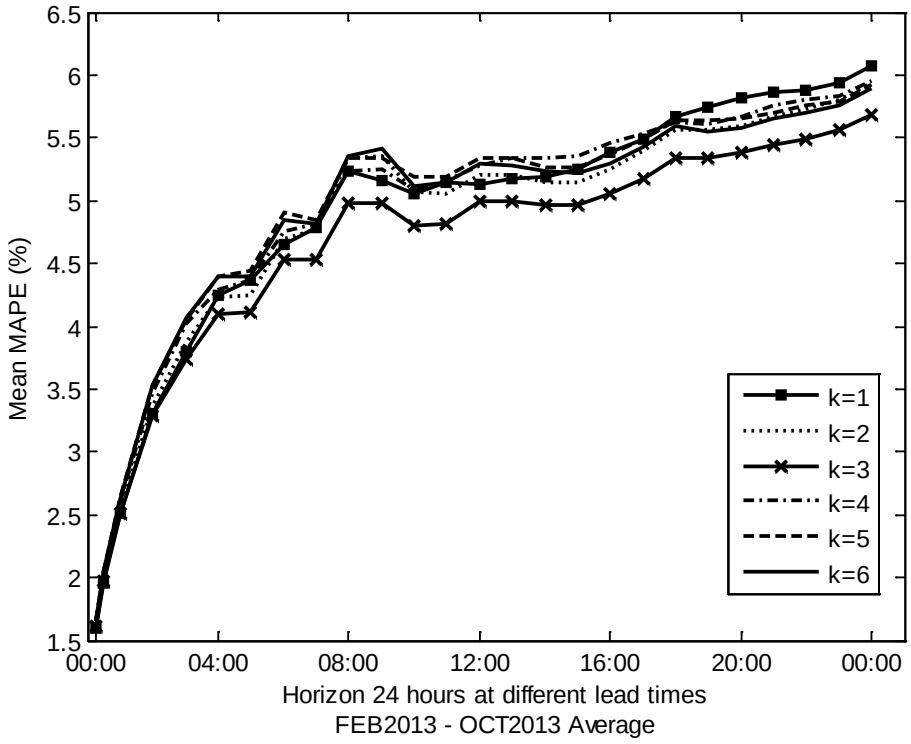

Figure 4-7 MAPE Results Plotted Against Lead Time for the 9-Month Out-of-Sample Period for Lead Times of 15min Ahead, 30min Ahead, 1h Ahead, 2h Ahead, …, 24h

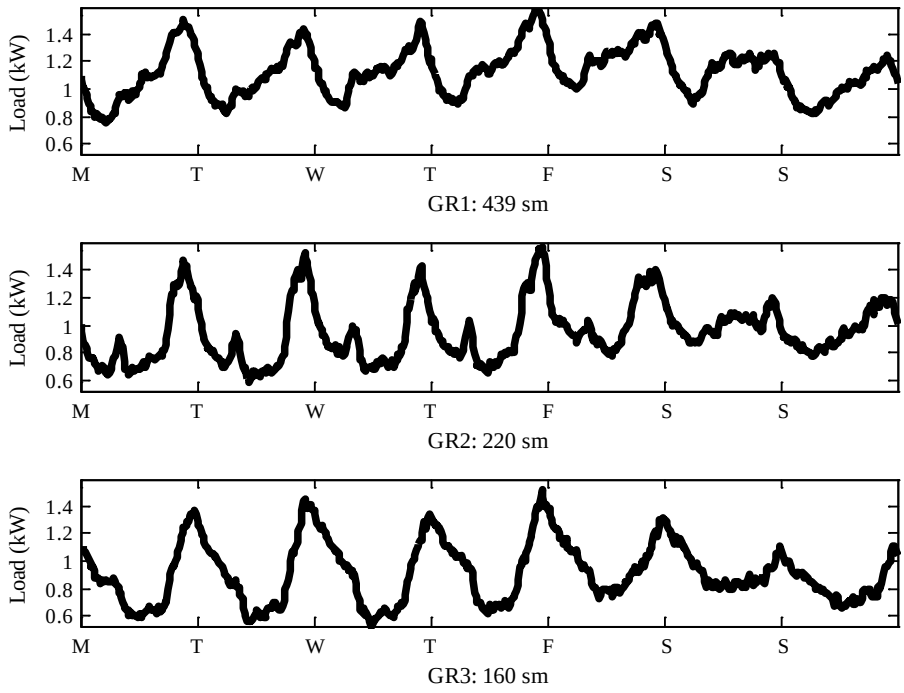

Figure 4-8 Load Profiles of the 3 Groups of Meters When k = 3, the Optimal Number of Clusters in Dataset 1

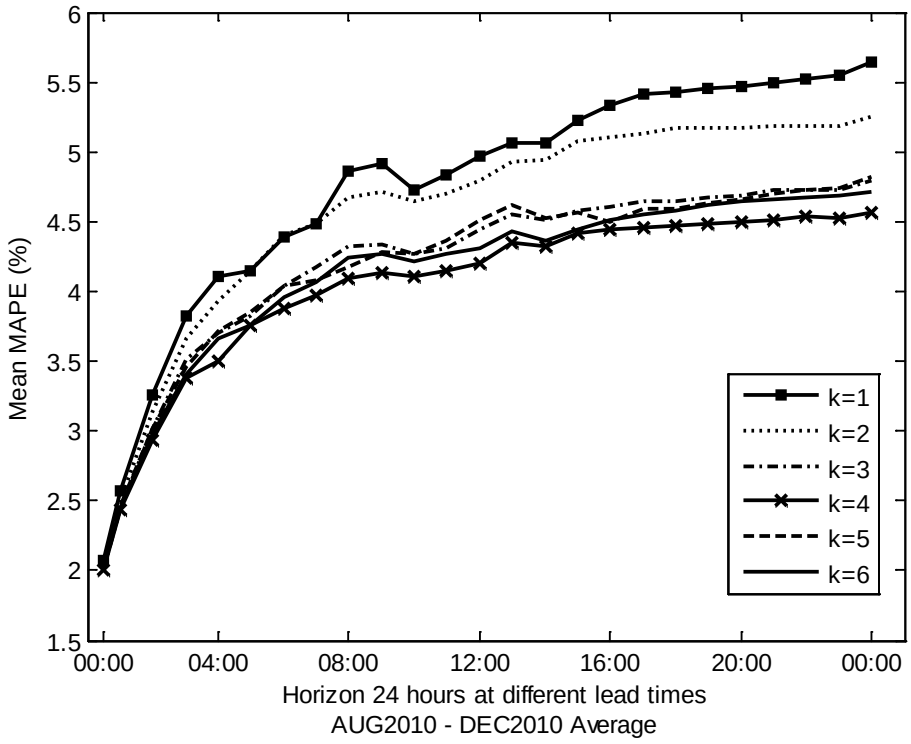

Figure 4-9 MAPE Results Plotted Against Lead Time for the 5-Month Out-of-Sample Period for Lead Times of 30min Ahead, 1h Ahead, 2h Ahead, …, 24h

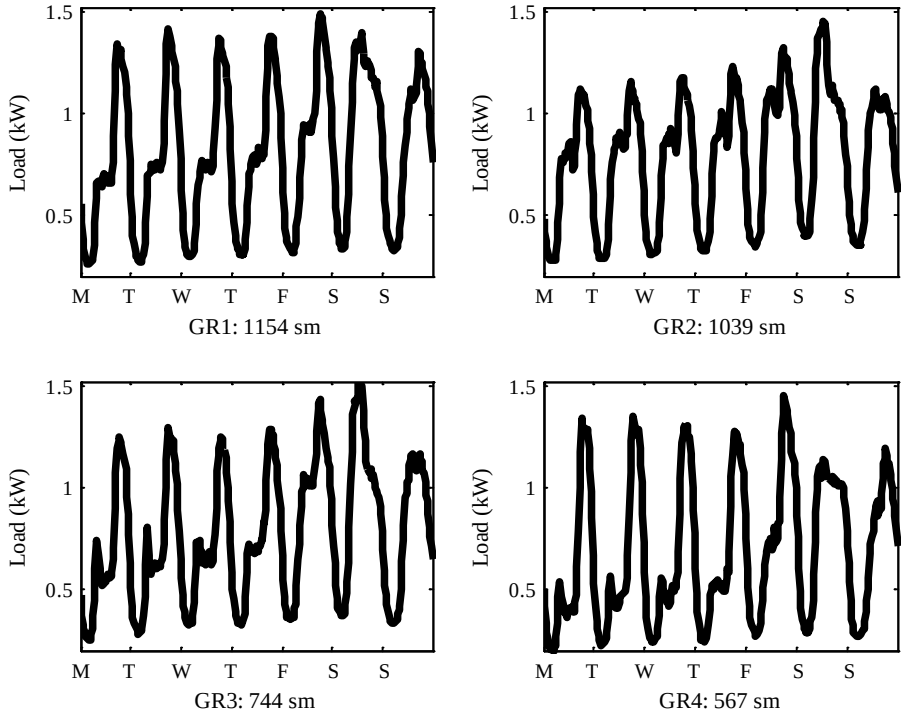

Figure 4-10 Load Profiles of the 4 Groups of Meters When k = 4, the Optimal Number of Clusters in Dataset 2

Chapter 5

Major Appliances Identification Considering ZIP Load Models – A Statistical Approach

The concept of non-intrusive load monitoring (NILM) is not new, but to our knowledge AMI data for this purpose is limited. Many researchers' approaches, as the technical literature indicates, are based on installing additional hardware (meter-extension) or sensors (at each appliance-ILM) besides the revenue meter with sampling rates ranging from 1 Hz to 100 MHz approx., and mainly developed in the laboratory, which, from the practical point of view, are not viable.

The purpose in this part of this dissertation is to investigate a viable solution for a realistic environment based on the AMI data and considering polynomial load models. Major Appliance Identification (MAI) considers a statistical approach based on predetermined databases, as a minimum, for identifying load components with emphasis on major appliances because of the sampling rate.

The polynomial model is one of the static load models widely accepted among utility industries. This model is commonly referred as the ZIP model, as it is composed of Constant Impedance-Current-Power components.

5.1 Databases Development for Major Appliances Identification

MAI program requires the use of predetermined data, as a minimum, for identifying load components with emphasis on major appliances. Careful research was done as part of this present work to determine a set of default databases. Load components models, typical rated power consumption of the load components, end-use profiles for typical average days of the load components, and typical time of use of the load components are the databases taken into consideration that MAI program uses to uniquely identify appliances' operation.

The subject of data sources will be dealt with in this section. A procedure for building a “database” for major appliances identification, as well as verifying its accuracy, would consist of the following steps:

- Collect available information on load components that exist in the utility’s power system distribution grid.
- Obtain any information is available on load components
    - ZIP load models
    - Typical rating values
    - Daily load curve shapes
    - Typical time of use
- Synthesize all this information
- Develop the databases

5.2 End-Use Load Categories and Components in the Utility Service Area

Load modeling is for all power system studies, as important as the rest of the power system models; however, most utilities treat their load as constant PQ in their simulations due to the complexity in obtaining accurate models to represent the load behavior. Although this approach is adequate for some applications, utility companies recognize that substantial changes in the nature of their supplied electric load have made it increasingly clear that a more appropriate load representation is needed. Even though the load is inherently random, unpredictable depending on lifestyle, weather, and many other uncontrollable factors, it is still possible to find a suitable model that can/will improve accuracy of simulations.

In [63] and [64], the authors have conducted extensive load survey at various residential and commercial customers at the utility company. In [63], it has been established load models for eighteen load components in order to provide an accurate

representation of the load under low voltage contingency conditions. According to [63], the field surveys helped to determine which types of equipment and appliances were in use at commercial and at residential sites in the utility service area. Those models were based on the existing equipment during the period in which they were developed. Although [63] provided a significant contribution towards improving the utility's understanding of their load components, it is necessary to update and/or include new products that have emerged into the market after the development of the original load models. In this way, in [64], a series of new surveys in the utility company, which explores the changes occurred in the load composition over time, is established.

These studies have laid the foundation for establishing End-Use Categories and End-Use Load Components for typical "commercial" and "residential" customers [64], as shown in Table 5-1.

5.2.1 End-Use Load Component Models: PQ-ZIP Database

MAI approach requires the specification of the End-Use Load Component Models which are designed to incorporate the major components affecting energy use in the Utility residential sector. The End-Use Load Component models consist of ZIP load models corresponding to the most representative residential loads powered by the Utility. It is designated as PQ-ZIP Database. PQ stands for P, the active power, and Q the reactive power.

To obtain the ZIP load models for the present study it has been taken into consideration important sources of data provided by the utility company in [63] and [64], coupled with in-house testing models following the procedure as described in Section 5.2.1.1 whose preliminary results were presented in [65]. The sole purpose of all of these is to get a solid ZIP load models database as possible based on the available information

for the load component characteristics that represent different makes, brands, and technologies found in the Utility's distribution system.

Table 5-1 Residential End-Use Categories and Load Components

End-Use Category	End-Use Load Component
Air Handling	Air conditioner
	Chillers
	Ventilators
	Fans
Compressor	Air compressor
	Industrial freezer
Pump	Hot & cold water circulation
	Chiller pump
	Fire pump
Lighting	Fluorescent
	Incandescent
	Halogen
	Compact fluorescent
Kitchen Appliances	Refrigerator
	Freezer
	Warmer
	Oven
	Microwave oven
Laundry Appliances	Washer
	Dryer
Electronics	Power supply
	Television
	Computer
	Peripheral
Elevators	Hydraulic
	Pneumatic
	Traction

5.2.1.1 ZIP load component models in-house testing

Substantial changes in the nature of the loads have occurred recently with the development of the new technologies. This makes clear the necessity to update the load model and provide an accurate representation of the new generation loads.

Mathematically, the ZIP model is represented as follows:

$$P = P_0 \left[Z_p \cdot \left(\frac{V}{V_0}\right)^2 + I_p \cdot \left(\frac{V}{V_0}\right) + P_p \right] \qquad (5.1)$$

$$Q = Q_0 \left[Z_q \cdot \left(\frac{V}{V_0}\right)^2 + I_q \cdot \left(\frac{V}{V_0}\right) + P_q \right] \qquad (5.2)$$

where P is the total active power, Q is the total reactive power, Z_i , I_i, P_i, $i = \{p, q\}$ are constant impedance, constant current, and constant power fractions, respectively, and they constitute the parameters to be determined. P_0 and Q_0 are the load active and reactive power respectively at rated voltage V_0.

5.2.1.1.1 Testing

The load component testing was executed in a controlled environment. The required power is obtained from the grid through an auto-transformer, and only voltage excursions are considered through a stepwise variation of the voltage in a slow ramp (voltage rampdown test) to develop the static load model. Starting at 130V, the voltage is decreased in steps of 5V until the device shuts down. The shut down voltage is referred as V_{off}. Even though the voltage of operation of some appliances is rated in a range (e.g. 110-120V), 120V and 60 Hz are defined as the nominal values. Standard laboratory instrumentation was used in these tests. The quantities monitored are voltage, current, real power, reactive power, and power factor recorded at a sampling rate of ½ second (due to logger equipment capabilities) for later analysis.

5.2.1.1.2 Data handling

All recorded data is in a raw format, so it is necessary to extract refined data from them. All raw data has been plotted as a means of identifying valid data. Therefore, the entire test sequence can be visually previewed, and the valid test data can be specified by a beginning and ending sample number. This treatment converts the test data to a usable form.

5.2.1.1.3 Determination of the ZIP coefficients

The determination of the ZIP coefficients is formulated as an optimization problem. Let L be the error to be minimized,

$$L = \sum_{i=1}^{N} L_i^2 = \sum_{i=1}^{N} (g(V_i) - g_i)^2 \qquad (5.3)$$

Here, g can represent P or Q, the ZIP load models. Therefore,

$$L = \sum_{i=1}^{N} \left(Z_p \cdot \left(\frac{V_i}{V_0}\right)^2 + I_p \cdot \left(\frac{V_i}{V_0}\right) + P_p - \frac{P_i}{P_0} \right)^2 \qquad (5.4)$$

V_i/V_0 and P_i/P_0 correspond to normalized values of voltage and power respectively, with respect to their nominal values; N is the number of samples.

The ZIP coefficients can be fitted to the measured data considering three cases:

1. No constraints; therefore, minimization of (5.4).
2. One constraint added: minimization of (5.4) subject to the sum of all coefficients should be equal to one (5.5). This constraint ensures that the load consumes the correct power at nominal voltage [66].

$$h(Z_p, I_p, P_p) = Z_p + I_p + P_p - 1 = 0 \qquad (5.5)$$

3. Two constraints added: minimization of (5.4) subject to the sum of all coefficients should be equal to one (5.5), and all coefficients should greater than or equal to zero (5.6).

$$Z_p, I_p, P_p \geq 0 \qquad (5.6)$$

For cases 1 and 2, it should be noticed that the coefficients can be negative values, making the model not physically based, as long as the model behavior matches the load characteristic. The reactive power is treated in a similar fashion to determine its ZIP coefficients.

5.2.1.1.4 Case examples for ZIP load model coefficients determination

The steady-state ZIP load models resulting from the foregoing procedure is presented here for a 55-inch LCD-TV and a 55-inch LED-TV as case examples. The curves of Active Power vs. Voltage (PV) and Reactive Power vs. Voltage (QV) of the measurements and the ZIP models are shown in Figure 5-1 and Figure 5-2. For voltages above 100 V (approximately), one may notice that the TVs load consumption are nearly constant power for both LCD and LED technologies.

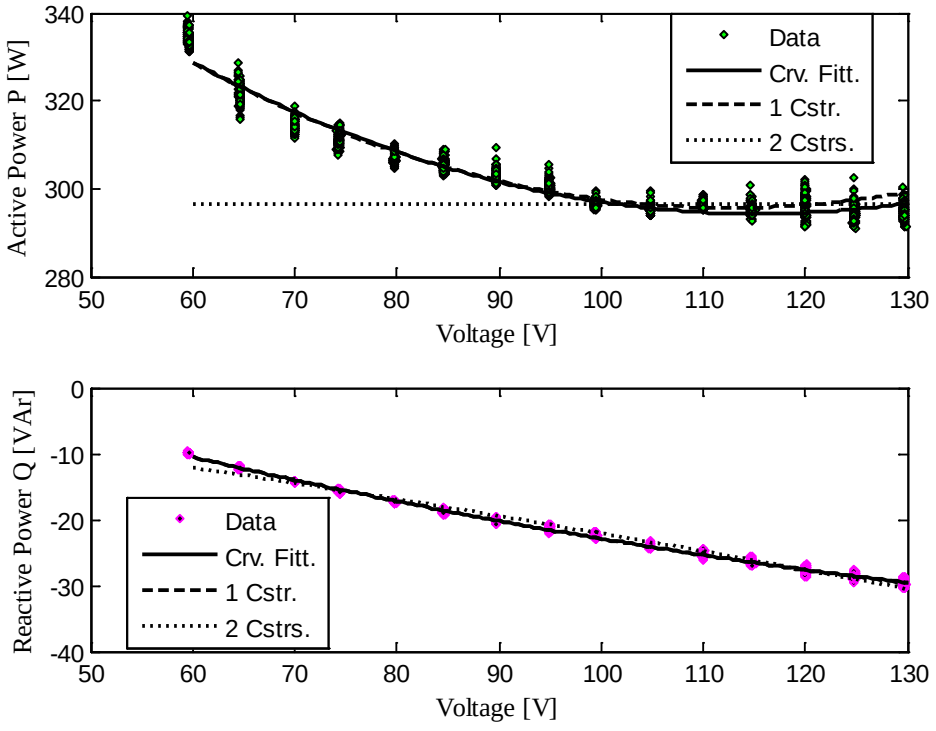

Figure 5-1 PV and QV Curves - 55" LCD Television

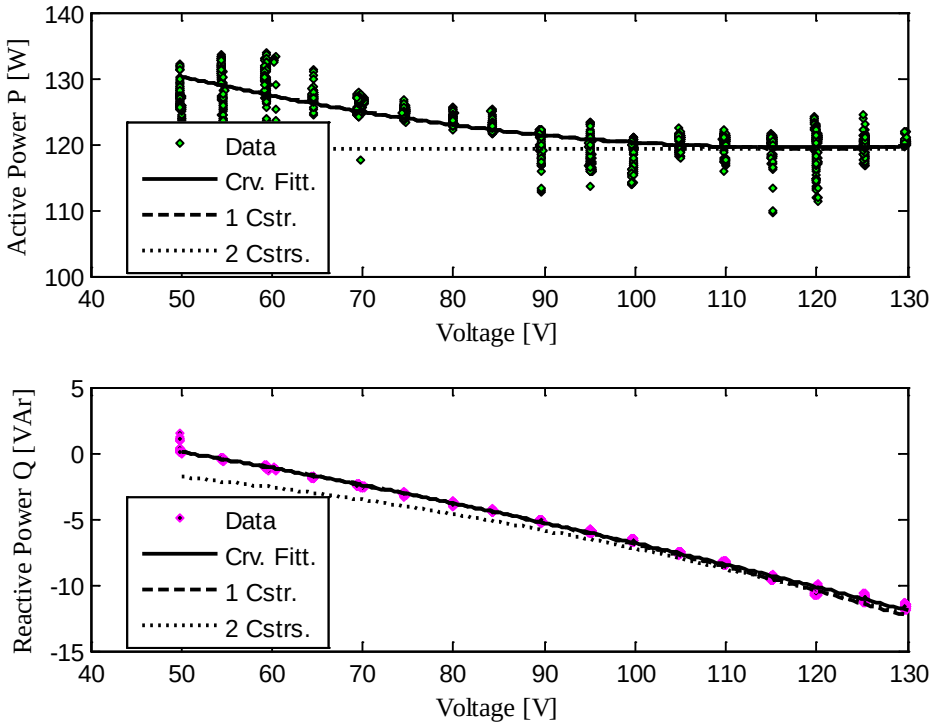

Figure 5-2 PV and QV Curves - 55" LED Television

Table 5-2 and Table 5-3 show the ZIP coefficients for the LCD-TV and the LED-TV respectively. The ZIP coefficients have been determined for the three cases N = No Constraints, 1 = One Constraint, and 2 = Two Constraints.

Table 5-2 ZIP Models for the 55-Inch LCD-TV

V_0 (V)	V_{off} (V)	Active Power P_0 (W)			Reactive Power Q_0 (VAr)		
120	50	296.29			-27.65		
Constraints		Z_p	I_p	P_p	Z_q	I_q	P_q
N		0.546	-1.051	1.499	-0.646	2.204	-0.560
1		0.580	-1.088	1.509	-0.637	2.194	-0.557
2		0.000	0.000	1.000	0.247	0.753	0.000

Table 5-3 ZIP Models for the 55-Inch LED-TV

V_0 (V)	V_{off} (V)	Active Power P_0 (W)			Reactive Power Q_0 (VAr)		
120	46	119.36			-10.44		
Constraints		Z_p	I_p	P_p	Z_q	I_q	P_q
N		0.276	-0.546	1.271	0.546	0.909	-0.482
1		0.271	-0.540	1.269	0.726	0.697	-0.423
2		0.000	0.000	1.000	1.000	0.000	0.000

5.2.1.2 PQ-ZIP database

The steady-state PQ ZIP load models database resulting from the foregoing procedure considering one constraint and references [63] and [64] are summarized in Table 5-4, Table 5-5, and Table 5-6. Appliances in blue background correspond to UTA developed models.

Table 5-4 PQ-ZIP Single Phase Load Models Database

No.	Device	S_0 (VA)	P_0 (W)	Q_0 (VAr)	pf	V_{off} (V)	V_0 (V)	Active Power			Reactive Power		
								Z_p	I_p	P_p	Z_q	I_q	P_q
1	No Candidate SP[6]	0	0	0	0	0	120	0	0	0	0	0	0
2	Air Conditioner #1	497.93	483.00	121.00	0.9700	100	120	1.74	-2.92	2.18	21.01	-36.57	16.56
3	Air Conditioner #2	529.96	513.00	133.00	0.9680	100	120	0.28	-0.13	0.85	9.51	-16.07	7.56
4	Laptop Charger	80.45	35.90	72.00	0.4462	100	121	-0.27	0.47	0.80	-0.36	1.22	0.14
5	Air Compressor 1ph	1215.80	1114.00	487.00	0.9163	100	120	0.75	0.38	-0.13	-0.92	3.23	-1.31
6	Fan #1	73.52	33.40	65.50	0.4543	100	121	-0.98	2.17	-0.19	-1.03	2.59	-0.56
7	Fan #2	312.46	295.00	103.00	0.9441	100	121	-0.01	1.29	-0.28	5.41	-8.25	3.84
8	Portable Fan	86.62	81.58	29.09	0.9419	50	120	0.48	0.70	-0.18	0.44	0.68	-0.12
9	Electronic Ballast (Advance)	59.21	59.00	4.94	0.9965	100	121	0.23	-0.51	1.28	10.16	-22.48	13.32
10	Electronic Ballast (GE)	62.86	61.50	13.00	0.9784	100	121	-0.02	1.24	-0.22	6.30	-9.71	4.41
11	Electronic Ballast (Universal)	61.72	61.50	5.23	0.9964	100	121	0.10	-0.24	1.14	4.43	-6.01	2.58
12	Magnetic Ballast	82.25	81.80	8.57	0.9946	100	121	-1.62	3.84	-1.22	35.76	-66.35	31.59
13	CFL Bulb #1	49.38	28.10	40.60	0.5691	100	120	1.21	-1.58	1.37	1.20	-1.27	1.07
14	CFL Bulb #2	41.71	23.30	34.60	0.5586	100	120	0.26	-0.22	0.96	0.54	-0.39	0.85
15	Incandescent Bulb	104.00	104.00	0.80	1.0000	100	121	0.45	0.65	-0.10	1.44	-1.09	0.65
16	Incandescent Eco Bulb	72.21	72.20	0.90	0.9999	100	120	0.47	0.61	-0.08	0.27	0.72	0.01
17	Refrigerator #1	131.71	120.00	54.30	0.9111	100	121	1.15	-1.76	1.61	6.97	-10.66	4.69
18	Refrigerator #2	159.15	91.80	130.00	0.5768	100	121	2.39	-3.84	2.45	2.51	-2.63	1.12
19	Refrigerator Top Mounted	146.97	143.47	17.11	0.9762	90	120	1.69	-3.40	2.71	16.43	-34.32	18.90
20	Halogen #1	109.00	109.00	0.91	1.0000	100	121	0.45	0.65	-0.10	0.78	-0.27	0.49
21	Halogen #2	95.71	95.70	1.01	0.9999	100	121	0.46	0.63	-0.09	-0.58	2.44	-0.86
22	Halogen #3	92.00	92.00	0.64	1.0000	100	121	0.46	0.64	-0.10	12.87	-22.43	10.56
23	LCD Television	209.06	208.00	-21.00	0.9949	100	121	0.11	-0.17	1.06	1.60	-1.72	1.12
24	LCD TV 55"	297.70	296.41	-27.65	0.9957	58	120	0.58	-1.09	1.51	-0.63	2.19	-0.55

[6] No Candidate SP is a reserved space to represent any Single Phase (SP) appliance whose characteristics do not match any of the known loads from 2 to 32.

Table 5-2—*Continued*

25	LED TV 55"	122.41	121.68	-13.38	0.9940	51	120	0.43	-0.81	1.39	0.41	1.02	-0.43
26	Microwave (GE) #1	1451.89	1405.00	366.00	0.9677	100	121	1.68	-2.51	1.83	60.21	-115.59	56.38
27	Microwave (Haier) #2	1446.39	1334.00	559.00	0.9223	100	121	1.31	-1.80	1.49	40.14	-71.30	32.16
28	Microwave #3	1252.20	1252.00	-22.13	0.9998	100	120	5.49	-9.30	4.81	-638.37	1088.14	-448.76
29	Microwave #4	929.69	922.60	114.62	0.9924	100	120	2.17	-2.84	1.67	101.07	-169.38	69.31
30	PC (Monitor & CPU)	208.85	117.00	173.00	0.5602	100	121	0.19	-0.26	1.07	0.07	0.48	0.45
31	DesktopPC-LCD24"	189.10	187.45	-24.87	0.9913	66	120	0.40	-0.72	1.32	2.34	-4.19	2.85
32	Vacuum Cleaner	897.91	869.00	226.00	0.9678	100	121	0.80	0.36	-0.16	3.38	-4.46	2.08

Table 5-5 PQ-ZIP Bi Phase Load Models Database

No.	Device	S_0 (VA)	P_0 (W)	Q_0 (VAr)	pf	V_{off} (V)	V_0 (V)	Active Power			Reactive Power		
								Z_p	I_p	P_p	Z_q	I_q	P_q
1	No Candidate BP[7]	0	0	0	0	0	240	0	0	0	0	0	0
2	Air Conditioner	1019	988	249.4	0.969	100	208	0	0	1	0	0	1

Table 5-6 PQ-ZIP Three Phase Load Models Database

No.	Device	S_0 (VA)	P_0 (W)	Q_0 (VAr)	pf	V_{off} (V)	V_0 (V)	Active Power			Reactive Power		
								Z_p	I_p	P_p	Z_q	I_q	P_q
1	No Candidate TP[8]	0	0	0	0	0	120	0	0	0	0	0	0
2	Air Compressor 3ph	1461.43	1175.00	869.00	0.8040	174	211	0.12	0.02	0.86	4.87	-7.64	3.77
3	Elevator Emulation (0.75HP)	995.05	802.00	589.00	0.8060	174	210	0.53	-0.99	1.46	1.96	-2.47	1.51
4	Elevator Emulation (1.0HP)	1150.25	988.00	589.00	0.8589	174	211	0.24	-0.44	1.20	5.71	-10.18	5.47
5	Elevator Emulation (1.5HP)	1723.81	1383.00	1029.00	0.8023	174	210	0.39	-0.70	1.31	3.80	-5.75	2.95

[7] No Candidate BP is a reserved space to represent any Bi Phase (BP) appliance whose characteristics do not match any of the known loads if they exist.
[8] No Candidate TP is a reserved space to represent any Three Phase (TP) appliance whose characteristics do not match any of the known loads from 2 to 5.

5.2.2 Typical Rated Power Consumption of Load Components: P_{range} Database

Important information of any electrical appliance can be found in its nameplate regarding the identifying name and the rating in volts and amperes or in volts and watts; or if the appliance is to be used on a specific frequency or frequencies [67]. The wattage of an appliance is stamped in most cases on the nameplate on the bottom of the back of the appliance. The wattage listed is the maximum power drawn by the appliance. It should be noted that the actual amount of power consumed depends on the setting used at any one time. Therefore, the wattage from an appliance's nameplate is just a reference value, and not its actual drawn power. If the wattage is not listed on the appliance, it can still be estimated by finding the current draw (in amperes) and multiplying that by the voltage (in volts) used by the load component.

P_{range} accounts for the variability in size and makes of the appliances in the PQ-ZIP database base case. Each load component listed before is associated with sets of reference active power values consumed by each of the appliance respectively. These reference values were obtained from different sources:

- Local department store nameplate recordings
- Datasheets from appliances' manufacturers
- Appliance Efficiency Database from the California Energy Commission [68]
- Energy Star Qualified Products [69]
- Estimating Appliance and Home Electronic Energy Use from the U.S. Department of Energy, Energy Efficiency & Renewable Energy [70]

5.2.3 Hourly Load Curve Shapes: Normalized End-Use Load Profile Database

Statistically averaged usage pattern of end-use appliances is utilized as means of determining which appliance is turned on most likely in the event of more than one appliance candidate is found during the identification process of MAI program.

This average usage pattern is what is referred to as “appliance load shape”, and it is a measure of the average electricity consumption of an appliance over the course of each hour on an average day. An “average” day could be an average annual day, an average summer day, or an average winter day [71]. All average hourly load profiles are normalized (dimensionless), and they represent an energy-weighted probability distribution of the load’s occurrence at any given hour [72]. For the purpose of this dissertation, the load shapes developed as part of the End-Use Load and Consumer Assessment Program are used. Two sources of information were explored to assist in the process of obtaining these curve shapes that are publicly available:

- The Building America Analysis Spreadsheets [73] by Building America House Simulation Protocols report [74]; for example Home Entertainment in Table 5-7, Figure 5-3.

Table 5-7 Building America - Home Entertainment Devices Hourly Load Profile

Winter (October - March) Normalized Energy Use Profile												
Hour	1	2	3	4	5	6	7	8	9	10	11	12
%	4.8	2.7	1.4	0.2	0.1	0.1	0.5	1.0	1.7	2.4	3.3	4.2
Hour	13	14	15	16	17	18	19	20	21	22	23	24
%	4.3	4.4	4.6	4.8	5.4	6.0	7.0	8.1	8.7	9.3	8.1	6.9
Summer (April - September) Normalized Energy Use Profile												
Hour	1	2	3	4	5	6	7	8	9	10	11	12
%	4.2	1.2	0.6	0.0	0.0	0.0	0.5	0.9	1.9	2.9	3.2	3.5
Hour	13	14	15	16	17	18	19	20	21	22	23	24
%	3.7	3.8	4.1	4.3	4.7	5.0	7.0	9.0	10.7	12.3	9.7	7.1

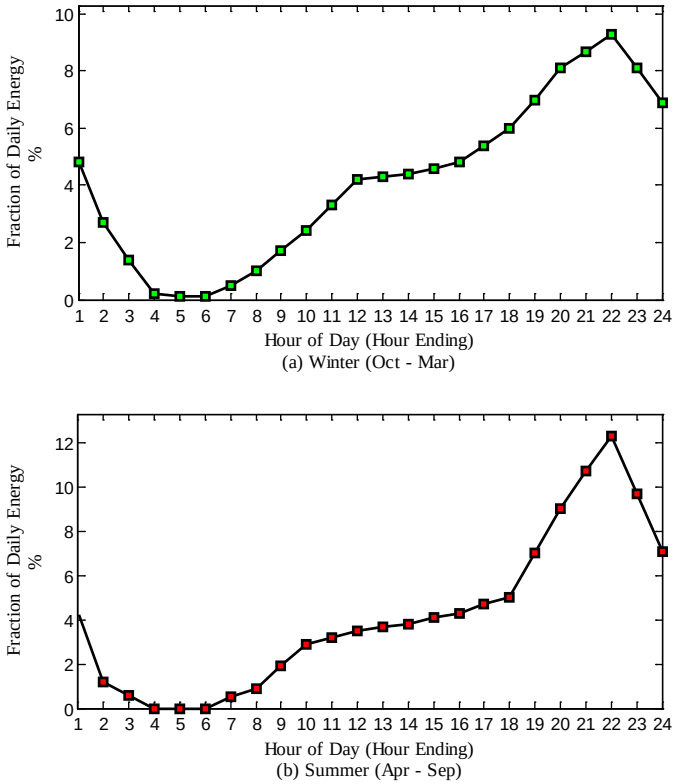

Figure 5-3 BA - Home Entertainment Devices Load Shape for an Average (a) Winter, and (b) Summer Day

- Repository files for GridLAB-D simulation software [75]; for instance Heating-Cooling (Air Conditioner) in Table 5-8, Figure 5-4.

Table 5-8 GridLAB-D Rep. Files - Air Conditioner Hourly Load Profile

Winter (October - March) Normalized Energy Use Profile												
Hour	1	2	3	4	5	6	7	8	9	10	11	12
Weekday (%)	3.2	3.4	3.6	3.8	4.1	5.1	6.2	6.2	5.6	4.9	4.3	3.9
Weekend (%)	3.3	3.4	3.6	3.8	4.1	4.8	5.5	6.2	6.3	5.7	4.9	4.3
Hour	13	14	15	16	17	18	19	20	21	22	23	24
Weekday (%)	3.7	3.5	3.4	3.6	4.1	4.4	4.3	4.1	4.0	3.9	3.5	3.2
Weekend (%)	3.9	3.7	3.6	3.6	3.8	3.9	3.8	3.8	3.9	3.8	3.5	3.2
Summer (April - September) Normalized Energy Use Profile												
Hour	1	2	3	4	5	6	7	8	9	10	11	12
Weekday (%)	2.1	1.8	1.6	1.4	1.4	1.6	2.3	2.8	3.0	3.3	3.7	3.9
Weekend (%)	2.1	1.7	1.6	1.4	1.4	1.6	2.1	2.8	3.3	3.5	3.8	4.2
Hour	13	14	15	16	17	18	19	20	21	22	23	24
Weekday (%)	4.4	5.1	5.8	6.6	7.5	8.0	8.2	7.5	6.3	5.1	4.0	2.8
Weekend (%)	4.7	5.2	5.9	6.6	7.3	7.8	8.0	7.3	6.1	4.9	3.8	2.8

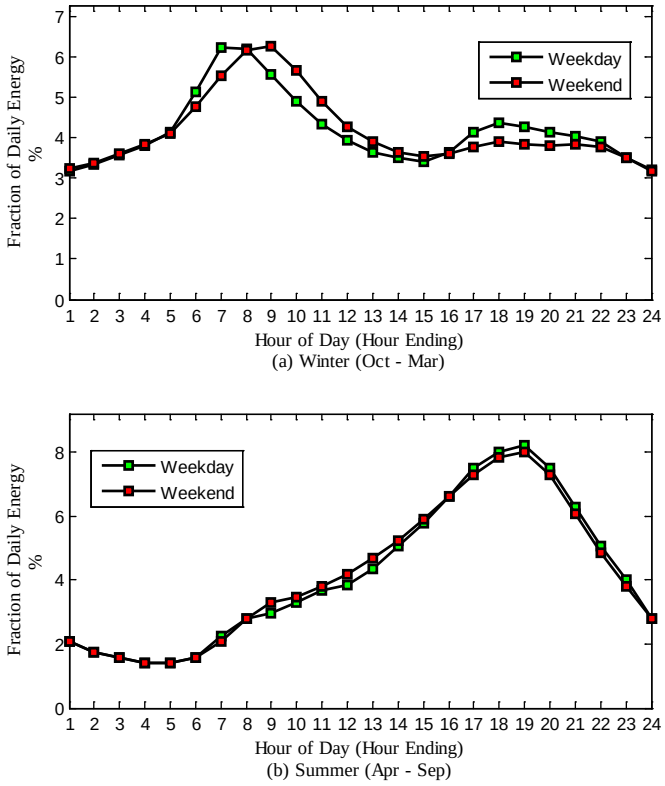

Figure 5-4 GLD - Air Conditioner Load Shape for an Average (a) Winter, and (b) Summer Weekday, and Weekend

5.2.3.1 EUNLP assignment to the end-use load components for MAI

After the End-Use Normalized Load Profiles have been introduced in the previous section, it is time to assign them to the End-Use Load Components. EUNLPs from the GridLAB-D Repository files are preferred whenever possible because of their weekdays and weekends distinction. EUNLP from Building America are utilized when there is no load component in EUNLP-GLD that can describe its corresponding ZIP load component.

Table 5-9, Table 5-10, and Table 5-11 provide the EUNLP assignment to the End-Use Load Components that are used in the MAI program for Single, Bi, and Three Phase appliances respectively.

5.2.4 Typical Interval Time of Use Load Components Database

Each load component has been associated with sets of typical interval Time of Use by each of the appliance respectively. Therefore, Typical Interval Time of Use can improve MAI program's ability to match characteristics to observed data and help explain load behavior.

In recent years, load disaggregation has drawn renewed interest from the research community, making it possible to provide publicly usage information from real homes consumption. This usage information has been processed to obtain the typical interval time of use as well as other available sources:

- The Reference Energy Disaggregation Data Set (Initial REDD Release, Version 1.0) [76]
- Building-Level Fully Labeled Electricity Disaggregation Data Set (BLUED) [77]
- The Building America Analysis Spreadsheets [73] by Building America House
- ResPoNSe: modeling the wide variability of residential energy consumption [78]

Table 5-9 End-Use Normalized Load Profile Assignment for Single Phase Loads

No.	Device	EUHLP-GLD	EUHLP-BA
1	No Candidate SP	-	-
2	Air Conditioner #1	Heating Cooling	-
3	Air Conditioner #2	Heating Cooling	-
4	Laptop Charger	-	Other MELs
5	Air Compressor 1ph	Freezer	-
6	Fan #1	-	Ceiling Fans
7	Fan #2	-	Ceiling Fans
8	Portable Fan	-	Ceiling Fans
9	Electronic Ballast (Advance)	Light Plugs	-
10	Electronic Ballast (GE)	Light Plugs	-
11	Electronic Ballast (Universal)	Light Plugs	-
12	Magnetic Ballast	Light Plugs	-
13	CFL Bulb #1	Light Plugs	-
14	CFL Bulb #2	Light Plugs	-
15	Incandescent Bulb	Light Plugs	-
16	Incandescent Eco Bulb	Light Plugs	-
17	Refrigerator #1	Refrigerator	-
18	Refrigerator #2	Refrigerator	-
19	Refrigerator Top Mounted	Refrigerator	-
20	Halogen #1	Light Plugs	-
21	Halogen #2	Light Plugs	-
22	Halogen #3	Light Plugs	-
23	LCD Television	-	Home Entertainment
24	LCD TV 55"	-	Home Entertainment
25	LED TV 55"	-	Home Entertainment
26	Microwave (GE) #1	Food Preparation	-
27	Microwave (Haier) #2	Food Preparation	-
28	Microwave #3	Food Preparation	-
29	Microwave #4	Food Preparation	-
30	PC (Monitor & CPU)	Other	-
31	DesktopPC-LCD24"	Other	-
32	Vacuum Cleaner	Other	-

Table 5-10 End-Use Normalized Load Profile Assignment for Bi Phase Loads

No.	Device	EUHLP-GLD	EUHLP-BA
1	No Candidate TP	-	-
2	Air Conditioner	Heating	-

Table 5-11 End-Use Normalized Load Profile Assignment for Three Phase Loads

No.	Device	EUHLP-GLD	EUHLP-BA
1	No Candidate TP	-	-
2	Air Compressor 3ph	Freezer	-
3	Elevator Emulation (0.75HP)	Other	-
4	Elevator Emulation (1.0HP)	Other	-
5	Elevator Emulation (1.5HP)	Other	-

5.2.5 Adaptive PQ-ZIP Database

Initially this database can be filled out with the appliances presented in Section 5.2.1 to establish a good starting point. Keep in mind that the appliance which is labeled “No Candidate” (NC) corresponds to a special tag reserved for any appliance that has not been named[9] yet, and gives flexibility to add a new element whenever a match of NCs is found during an ON-OFF operation, and it is established a known relationship with the stored appliances in the PQ-ZIP Database during the appliance identification process. Once the identification is possible through the algorithm, a new element is added into this database. This happens every time a new appliance is matched, and identified.

Therefore, this database adapts whenever a new appliance appears.

$$P_{0s} \quad | \quad Q_{0s} \quad | \quad n_{lc}$$

where,

P_{0s} is the tailored active power at rated voltage V_0

Q_{0s} is the tailored reactive power at rated voltage V_0

n_{lc} is the index that identifies uniquely to a load component, and it is used to extract the ZIP load coefficients from the PQ-ZIP load database.

The adaptive PQ-ZIP Database has three files with each one storing a different set of appliances. The first file stores adaptive PQ-ZIP load components for Single-Phase loads, the

[9] Named and identified means similar concepts. By named, it indicates that the appliance has been given a name, e.g. air conditioner, and by identified means that the appliance is identified to be a known appliance, e.g. the air conditioner. Therefore, they are used indistinctively.

second file stores adaptive PQ-ZIP load components for Bi-Phase[10] loads, and the third file stores adaptive PQ-ZIP load components for Three-Phase loads. All of them have reserve index $n_{lc} = 1$ to denote a No Candidate.

5.3 Major Appliances Identification Algorithm

The software starts with the specification of the input data file to be analyzed. No manual training stage is necessary because the software will learn automatically during the identification process. Every time a new appliance appears and is recognized by MAI program, it will be catalogued and stored in the adaptive PQ-ZIP database.

5.3.1 Data Reading

Database initialization is performed where MAI program will read and load:

1. ZIP load models
2. Active power range
3. Normalized end-use load hourly profile
4. Known CT ratios
5. Adaptive ZIP load models

MAI algorithm is intended to work in real-time, so the algorithm is implemented in a recursive form by passing once through the data. During the algorithm design stage, the data are read as a block per day for a particular customer as they are available. Refer to Chapter 2 for detailed information about smart meter readings.

5.3.2 Edge Detection Based on Current and Phase Identification: 3-Phase 2-Level Edge Detection Algorithm

Load disaggregation platform will detect when there is an actual appliance being switched (on/off) by identifying changes in steady-state level. A triggering threshold scheme is used as means of detecting positive and negative edges. Careful observation and testing of the

[10] By bi-phase load it is understood that it corresponds to a load that is connected to two phases out of the three phase system.

available channels that provide edge information has made possible to determine that rather than using power measurements on the edge detection, current measurements can be used.

Current measurements have proven to provide stable changes because almost all the transitions (down-up and up-down steps) occur within one step –or one minute. Whereas the active power measurements were seen to ramp up/down, in some cases, in 2 steps or even more for the same event; this is an undesirable condition because the objective is to be able to identify appliance operations in real time. Moreover, the first measurement after a transition occurs is uncertain because on/off operation can happen at anytime within 1 minute. Therefore, this condition is skipped by taking the next measurement, as far as the detection is concerned, where a steady period is assumed to be reached after a maximum time of 1-minute following the transition. In other words, the identification is possible with 1-minute delay. More than 2 steps to reach steady state values is equivalent to say that the appliance takes 3 minutes and beyond to become in steady state operation. It can happen of course, and the current should follow this behavior. However, this condition is beyond MAI's capability as it will be explained later.

Alternatively, a change in voltage following an appliance operation was examined; however this change is small, can be masked by noise, and is not a real appliance operation.

The 3-phase 2-level edge detection algorithm is based on the current measurements. From a practical point of view, to detect a step change in the current value, a threshold thr_I is defined. All outputs below this threshold are discarded because from the project's perspective, the utility is interested in major appliances operation, and with steady state measurements at 1-minute sampling interval it is not expected to be able to identify small appliances. This threshold could be meter-specific, and should be above noise levels.

Firstly, flag statuses are initialized, and they are intended to help in identifying an edge nature. Let k be the present time, in minutes or time index, and I_k or $I(k)$ the present

measurement of the current I at time k, in amps. The algorithm compares two consecutive current measurements, at each corresponding phase A, B, and C,

$$(\Delta I)_1 = I(k) - I(k-1) \qquad (5.7)$$

If there is a significant change, i.e. any current variation of 5 A or more (5 A is the current threshold thr_I), a flag status is set to 1 indicating that there is an event at time k and it is ready to check the $k-2$ variation; otherwise there is no any event, and that particular phase is in a "Steady State" condition. All the current measurements and variations get to this point, and it is called level one. Once the flag status is activated, then the current variation $(\Delta I)_2$ for $k-2$ is calculated,

$$(\Delta I)_2 = I(k) - I(k-2) \qquad (5.8)$$

$(\Delta I)_1$ and $(\Delta I)_2$ are then compared against $\pm thr_I$ to set flag indicators up and down for level one, $(fl_{level1})_{up1}$, $(fl_{level1})_{up2}$, $(fl_{level1})_{down1}$, and $(fl_{level1})_{down2}$ accordingly:

- **If** $(\Delta I)_1 > thr_I$, $(fl_{level1})_{up1} = 1$; **Else**, $(fl_{level1})_{up1} = 0$, **End if**
- **If** $(\Delta I)_1 < -thr_I$, $(fl_{level1})_{down1} = 1$; **Else**, $(fl_{level1})_{down1} = 0$, **End if**
- **If** $(\Delta I)_2 > thr_I$, $(fl_{level1})_{up2} = 1$; **Else**, $(fl_{level1})_{up2} = 0$, **End if**
- **If** $(\Delta I)_2 < -thr_I$, $(fl_{level1})_{down2} = 1$; **Else**, $(fl_{level1})_{down2} = 0$, **End if**

After setting the flag indicators up and down for level one, the status level is changed to 2. Then, the edge detection algorithm advances to time $k' = k + 1$, where the whole procedure is repeated without reinitializing any flag, $(\Delta I)'_1$ and $(\Delta I)'_2$ are calculated for the new measurements, and it is time to compare them against $\pm thr_I$ to set flag indicators up and down for level two, $(fl_{level2})_{up1}$, $(fl_{level2})_{up2}$, $(fl_{level2})_{down1}$, and $(fl_{level2})_{down2}$ with a logic similar as before:

- **If** $(\Delta I)'_1 > thr_I$, $(fl_{level2})_{up1} = 1$; **Else**, $(fl_{level2})_{up1} = 0$, **End if**
- **If** $(\Delta I)'_1 < -thr_I$, $(fl_{level2})_{down1} = 1$; **Else**, $(fl_{level2})_{down1} = 0$, **End if**
- **If** $(\Delta I)'_2 > thr_I$, $(fl_{level2})_{up2} = 1$; **Else**, $(fl_{level2})_{up2} = 0$, **End if**
- **If** $(\Delta I)'_2 < -thr_I$, $(fl_{level2})_{down2} = 1$; **Else**, $(fl_{level2})_{down2} = 0$, **End if**

When the status level flag is 2, it is time to decide the nature of the edge. The edge nature is decided based on the flag indicators up and down that were assigned for levels one and two, and they have been grouped in a matrix designated as $\mathbf{F}_{\mathrm{LVL1\&2}}$ for convenience,

$$\mathbf{F}_{\mathrm{LVL1\&2}} = \begin{bmatrix} (fl_{level1})_{up1} & (fl_{level2})_{up1} & (fl_{level1})_{down1} & (fl_{level2})_{down1} \\ (fl_{level1})_{up2} & (fl_{level2})_{up2} & (fl_{level1})_{down2} & (fl_{level2})_{down2} \end{bmatrix}$$

There are eleven cases that have been considered to determine the edge nature, as they are shown in Table 5-12. Here, edge nature numbers 2, 5, 7, 8, 9, 10 and 11 indicate fast transitions up and down (ON) or down and up (OFF), and they do not provide meaningful information for appliance identification. Thus, edge nature 1, 3, 4, and 6 indicate that there is an operation (turn on, or turn off) of an appliance. If the transition meets one of the four meaningful edge nature conditions, a status flag for edge detection will be set to one at its corresponding phase A, B, and/or C. What follows next is to calculate the variation in power, active ΔW and reactive ΔVAr, that occur during this transition.

The variation in active and reactive power is no more than the difference between its k and $k-2$ values.

$$\Delta W(k) = W(k) - W(k-2) \qquad (5.9)$$

$$\Delta VAr(k-2) = VAr(k) - VAr(k-2) \qquad (5.10)$$

It should be noticed that the calculations in all equations in this Section 5.3.2 are performed at each phase A, B, and C. In the notation showed above the letters A, B, and C were dropped for convenience. Therefore, for ΔW, for instance, there exist ΔW_A, ΔW_B, and ΔW_C, and so forth for all the expressions.

Table 5-12 Edge Nature Definitions

Edge Nature No.	Edge Nature Name	$\mathbf{F}_{\mathrm{LVL1\&2}}$
1	Upward	$\begin{bmatrix}1 & 0 & 0 & 0\\ 1 & 1 & 0 & 0\end{bmatrix}$
2	Up/Down 1 minute	$\begin{bmatrix}1 & 0 & 0 & 1\\ 1 & 0 & 0 & 0\end{bmatrix}$
3	Ramp Up 2 minutes	$\begin{bmatrix}1 & 1 & 0 & 0\\ 1 & 1 & 0 & 0\end{bmatrix}$
4	Downward	$\begin{bmatrix}0 & 0 & 1 & 0\\ 0 & 0 & 1 & 1\end{bmatrix}$
5	Down/Up 1 minute	$\begin{bmatrix}0 & 1 & 1 & 0\\ 0 & 0 & 1 & 0\end{bmatrix}$
6	Ramp Down 2 minutes	$\begin{bmatrix}0 & 0 & 1 & 1\\ 0 & 0 & 1 & 1\end{bmatrix}$
7	Up half down	$\begin{bmatrix}1 & 0 & 0 & 0\\ 1 & 0 & 0 & 0\end{bmatrix}$
8	Down half up	$\begin{bmatrix}0 & 0 & 1 & 0\\ 0 & 0 & 1 & 0\end{bmatrix}$
9	Half down up half	$\begin{bmatrix}1 & 0 & 0 & 0\\ 0 & 0 & 0 & 0\end{bmatrix}$
10	Half up down half	$\begin{bmatrix}0 & 0 & 1 & 0\\ 0 & 0 & 0 & 0\end{bmatrix}$
11	Tick shape up	$\begin{bmatrix}1 & 0 & 0 & 0\\ 0 & 1 & 0 & 0\end{bmatrix}$

5.3.3 i-Phase Load Identification

Once the edge detection has been performed for all three phases A, B, and C, all the events were classified according to their edge nature, but it only tells us that something was turned ON or OFF indistinctly at each phase and independently. Therefore, i-Phase Load Identification determines whether the transitions correspond to a Single-Phase, Bi-Phase, or Three-Phase Load,

$$i \in \{Single, Bi, Three\}$$

A transition as a result of the edge detection can happen in either of three phases, in two phases, or in all three phases. A key value is then specified to indicate in which phase the transition exist, as shown in Table 5-13. Ones indicate the existence of a transition.

Table 5-13 Transition Key Value Designation

A	B	C	key
1	0	0	1
0	1	0	2
0	0	1	3
1	1	0	4
0	1	1	5
1	0	1	6
1	1	1	7

If key is either 1, 2 or 3, then the operation comes from a Single-Phase Load; likewise, it key is either 4, 5, or 6, then the operation is a Bi-Phase Load, or Single-Phase Loads at each one of the two phases. Finally, it is easy to notice what will happen if key is equal to 7.

In any key case, the appropriate active and reactive measured power, P_{msr}, and Q_{msr}, which would correspond to the power consumed by the load component are derived from the stored values in ΔW and ΔVAr, Table 5-14.

Table 5-14 Possible P_{msr} and Q_{msr} Consumed by the Load Component

key	P_{msr}	Q_{msr}
1	ΔW_A	ΔVAr_A
2	ΔW_B	ΔVAr_B
3	ΔW_C	ΔVAr_C
4	$\Delta W_A + \Delta W_B$	$\Delta VAr_A + \Delta VAr_B$
5	$\Delta W_B + \Delta W_C$	$\Delta VAr_B + \Delta VAr_C$
6	$\Delta W_C + \Delta W_A$	$\Delta VAr_C + \Delta VAr_A$
7	$\Delta W_A + \Delta W_B + \Delta W_C$	$\Delta VAr_A + \Delta VAr_B + \Delta VAr_C$

In the same way, in any key case, the appropriate values of the active and reactive power P_{ZIP} and Q_ZIP are calculated. P_ZIP and Q_ZIP are the active and reactive power, respectively, that are obtained from the ZIP load models tailored for a particular meter or group of meters, and stored in the Adaptive PQ-ZIP Database:

$$P_{ZIP}(V_{msr}) = P_{0s} \cdot \left[Z_P \cdot \left(\frac{V_{msr}}{V_0}\right)^2 + I_P \cdot \left(\frac{V_{msr}}{V_0}\right) + P_P \right] \qquad (5.11)$$

$$Q_{ZIP}(V_{msr}) = Q_{0s} \cdot \left[Z_Q \cdot \left(\frac{V_{msr}}{V_0} \right)^2 + I_Q \cdot \left(\frac{V_{msr}}{V_0} \right) + P_Q \right] \qquad (5.12)$$

where all the variables have been defined in its respective section, and only V_{msr} needs to be defined as the voltage at time k.

5.3.4 Measured and Calculated Values Comparison for Active and Reactive Powers

P_{msr} and Q_{msr} from a meaningful transition that occurs at time k have been obtained, as well as P_{ZIP} and Q_{ZIP} for all the possible known appliances from the Adaptive PQ-ZIP Database. It is now possible to compare them and determine whether the meaningful transition matches a possible candidate from the known appliances. Thus, it only compares the active/reactive power measured during a meaningful transition against the known appliances stored in the Adaptive PQ-ZIP Database. It should be kept in mind that this database has the first appliance labeled as "No Candidate" with $n_{lc} = 1$, while the others are the known appliances with $n_{lc} > 1$.

5.3.5 Tracking Transition Behavior and Match On-Off Operations

Every time a meaningful transition is determined to be an appliance changing state from OFF to ON (turning on); an ON-event is recorded. The information recorded in the ON-event is time-location of this event, and possible candidate index as it was obtained from Section 5.3.4. No further action is performed, just save the ON-event.

Once the meaningful transition is determined to be an appliance changing state from ON to OFF (turning off); it is time to match this OFF-event with its ON-event if exists.

The pairwise matching for an ON-OFF operation is basically based on what is turned off must have been turned on before. At the moment when an OFF-event exists, the appliance responsible for this operation is looked up on a pool of possible candidates that have been turn on already (ON-event possible candidates). If nothing has been turned on before, an error message is displayed indicating this matter.

To improve the accuracy on the matching procedure the Time of Use is taken into account. This means that matched ON/OFF appliances should be within their typical Time of Use, as described in Section 5.2.4.

5.3.5.1 If a match is found within the adaptive PQ-ZIP load models database

This means that an ON or OFF event is paired with a known appliance in the active database, and therefore, a proper index from Table 5-4, Table 5-5 or Table 5-6, if the load is single, bi, or three phase respectively.

5.3.5.2 If a match is found with "No Candidate" tag appliances

During the process described in Section 5.3.4, either of the ON or OFF events associated with the operation of an appliance can be labeled as NC[11]. First of all, if two NC appliances are matched together, a message is displayed indicating that a "Match is found for appliance # 1" followed by the ON and OFF locations. This match is saved in an output variable OutV,

$$OutV = [n_{lc} \quad t_0 \quad k \quad \Delta W(t_0) \quad \Delta W(k) \quad \Delta VAr(t_0) \quad \Delta VAr(k) \quad V_{msr}(t_0) \quad V_{msr}(k)]$$

So far, MAI has determined that there is an appliance that has been turned on at a certain location or time t_0, and it has been turned off at time $t_1 = k,\ t_0 < t_1$. However, this appliance is still unknown; therefore, when the match is made of NC appliances, the action will be to assign a suitable load model, or in other words, to find:

$$P_{0s} \quad | \quad Q_{0s} \quad | \quad n_{lc}$$

that represent most likely this appliance, and identify which kind of load component is, e.g. refrigerator, air conditioner, or other appliance with known signature given by the ZIP load model. Then it can be saved in the adaptive PQ-ZIP database. This procedure is called MAI Learning Algorithm.

5.3.5.2.1 MAI learning algorithm

Major Appliance Identification Learning Algorithm will "learn" which appliances a particular costumer (meter) has, based on an eight-step process:

1. Get the candidate measured data from the output variable $OutV$

[11] "No Candidate" appliance has the index number 1.

$OutV$ stores important information regarding an appliance's ON-OFF operation. Number 1 is the index that represents an appliance without any possible candidate, and this is the one that is going to be identified or named. From $OutV$'s information the following data can be derived,

$$P_1 = \frac{1}{2} \cdot \left(\Delta W(t_0) - \Delta W(k)\right) \quad (5.13)$$

$$Q_1 = \frac{1}{2} \cdot \left(\Delta VAr(t_0) - \Delta VAr(k)\right) \quad (5.14)$$

$$V_1 = \frac{1}{2} \cdot \left(V_{msr}(t_0) + V_{msr}(k)\right) \quad (5.15)$$

$$t_{1_{up}} = t_0 \quad (5.16)$$

$$t_{1_{down}} = k \quad (5.17)$$

2. Calculate the displacement power factor pf_1 from the measured data

   First, calculate the apparent power,

$$S_1 = \sqrt{(P_1)^2 + (Q_1)^2} \quad (5.18)$$

   then, calculate,

$$pf_1 = \frac{P_1}{S_1} \quad (5.19)$$

3. Calculate the displacement power factor PF_{ZIP} from the PQ-ZIP load models database

   PF_{ZIP} corresponds to the displacement power factor calculated from the ZIP load models. It is a vector of N elements, where N is the number of appliances in the PQ-ZIP database. It is understood that those values are obtained at V_1.

$$P(V_1) = P_0 \cdot \left[Z_P \cdot \left(\frac{V_1}{V_0}\right)^2 + I_P \cdot \left(\frac{V_1}{V_0}\right) + P_P\right] \quad (5.20)$$

$$Q(V_1) = Q_0 \cdot \left[Z_Q \cdot \left(\frac{V_1}{V_0}\right)^2 + I_Q \cdot \left(\frac{V_1}{V_0}\right) + P_Q\right] \quad (5.21)$$

$$S = \sqrt{P^2 + Q^2} \quad (5.22)$$

$$PF_{ZIP} = \frac{P}{S} \quad (5.23)$$

4. Compare pf_1 against PF_{ZIP}

   The power factor is chosen as signature which provides the necessary information about the electrical behavior of an individual appliance during steady-state operation. The assumption is that different loads of interest exhibit a unique power factor, and it accounts for the variability of the appliances size of the same kind. For instance, let's pick the LED TV which can be found in 32", 40", 55", and so on, and the power consumption will increase when the size increases. However, they all may share the same active and reactive power characteristic $P_C(V)$ and $Q_C(V)$ respectively but different P_0 and Q_0,

$$P_C(V) = Z_P \cdot \left(\frac{V_1}{V_0}\right)^2 + I_P \cdot \left(\frac{V_1}{V_0}\right) + P_P \qquad (5.24)$$

$$Q_C(V) = Z_Q \cdot \left(\frac{V_1}{V_0}\right)^2 + I_Q \cdot \left(\frac{V_1}{V_0}\right) + P_Q \qquad (5.25)$$

   A threshold thr_{pf} can be defined to make it easier the comparison of pf_1 against PF_{ZIP}. A binary list of the known appliances filled out with 1s when the pf_1 is close to PF_{ZIP}, otherwise 0s.

5. Determine if $P_{min} < P_1 < P_{max}$

   P_{min} and P_{max} define a range of possible values in which P_1 could exist. This range $[P_{min}, P_{max}]$ is taken from P_{range} Database as defined in Section 5.2.2. For example, let's continue talking about the LED TV. It is impractical (almost impossible) to get (test) the ZIP model for all TVs in the market (different sizes, different brands), but the most representative ones can be tested and then classified by the technology they use. Thus, one can test TVs such as LED, LCD, CTR, Plasma, and get their ZIP load models for these specific technologies to include them as benchmarks in the PQ-ZIP database. Keep in mind that P_1 can be overlapped in more than one type of load components.

6. Determine the time interval from $t_{1_{up}}$ to $t_{1_{down}}$ that the unknown appliance was ON

This time interval, or Time of Use (ToU) measured in minutes, will be used to test whether the unknown appliance has a time of use within the Typical Interval Time of Use of the possible known appliance candidates.

7. Determine which appliance is most likely to be ON according to the hourly load curve shapes

   Use $t_{1_{up}}$ and $t_{1_{down}}$ to determine the time of the day the appliance is ON, then estimate the probability of usage (PoU), from the Normalized End-Use Load Profile Database in Section 5.2.3, of the load models in the PQ-ZIP Database.

8. Finally, appliance naming

   Let's consider,

$$A = \{Appliance\ candidates\ found\ in\ step\ 4\colon pf_1 \in pf_{ZIP}\}$$

$$B = \{Appliance\ candidates\ found\ in\ step\ 5\colon P_1 \in [P_{min}\ ,\ P_{max}]\ \}$$

$$C = \{Appliance\ candidates\ found\ in\ step\ 6\colon ToU\}$$

   Then, the scheme represented in Figure 5-5 is followed to define if the "No Candidate" appliance can be identified, or there is definitely not a suitable candidate, and therefore "unidentifiable."

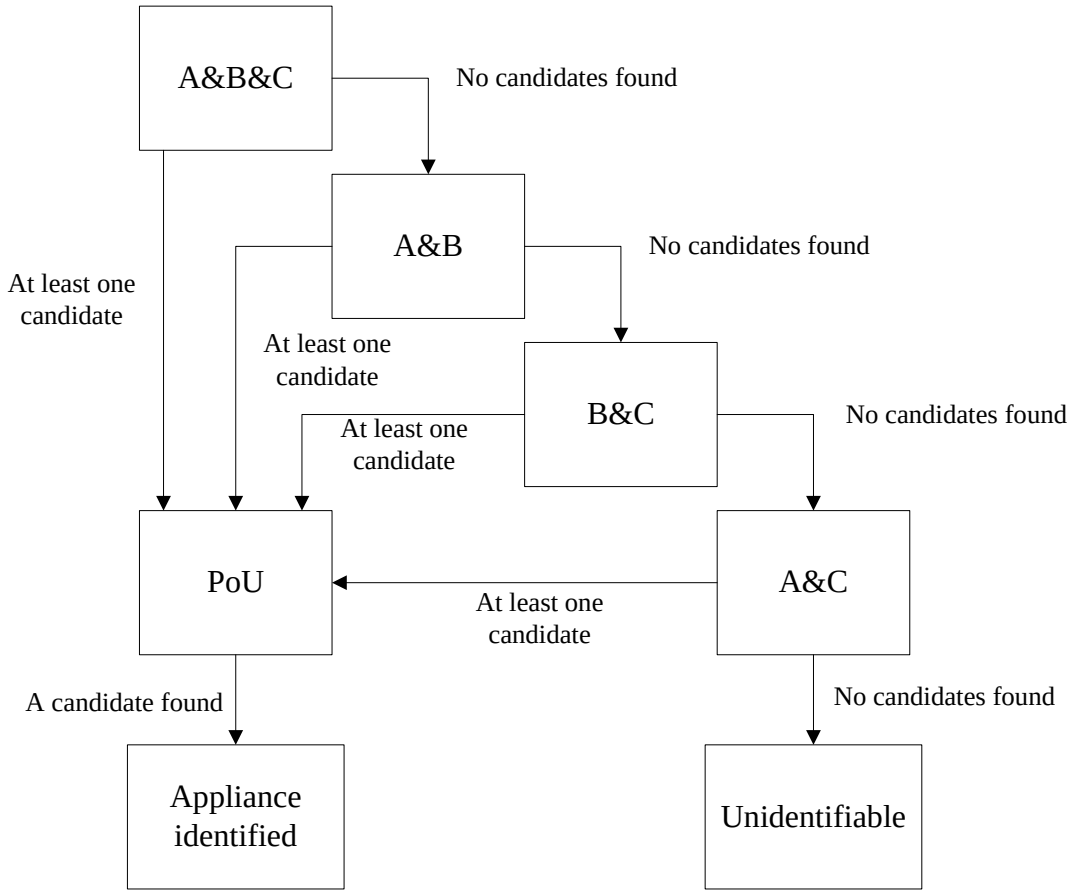

Figure 5-5 Decision Making Scheme to Determine if the "No Candidate" Appliance is Identifiable or Unidentifiable

Notice that if there is more than one possible candidate is found after $A\&B\&C$, $A\&B$, $B\&C$ and $A\&C$, PoU is used to rank them according to which known appliance is most likely to be operating first at that particular time of the day. Then, that one is the chosen one.

Chapter 6

Conclusions and Future Work Directions

Smart grid network paradigm relies on the exploitation of smart meter data to improve customer experience, utility operations, and advanced power management. In this dissertation, algorithms to utilize data collected from the AMI system of a utility company that can benefit directly to both, utility and customers have been conceptualized and developed. Applications of smart meter data can help the power utility company to improve the performance of its grid, and also improve its customers' experience.

A robust data reading and preprocessing tool has been developed that can handle large Smart Meter data files, and provide clean data ready for analysis. Nevertheless, there is still room for improvement and new algorithms will need to be proposed to validate smart meter readings and properly address missing values and outliers at individual customers.

Accurate customer daily load profiling has been developed for load estimation and network demand reconciliation to improve the efficiency and security of the utility grid. Load profiles were constructed based on customers' stratification information, and based on customers' consumption behavior similarities. These two approaches work well for smart metered customers, and future work should include assigning the load pattern inferred from the smart meters to the traditional meters' loads.

Real-world AMI data has been utilized to enhance the performance of load forecasting, which impacts operating practices and planning decisions to build, lease, or sell generation and transmission assets and the decisions to purchase or sell power at the wholesale level. The application of clustering to determine groups of customers considering load consumption similarities was demonstrated as an aid to improve the performance of load forecasting at the system level. It was shown how the load data at the household level from smart meters can be used to improve the load forecasting of the entire system by combining the forecasts from each group. While this characteristic was exploited for STLF based on Neural Networks, more work

should be performed to explore different techniques for STLF including probabilistic load forecasting techniques, and expand it to the study of MTLF, and LTLF.

A nonintrusive load monitoring (NILM) method was investigated for discerning individual appliances from a residential customer based on the AMI data. A comprehensive algorithm for Major Appliance Identification was developed, and typical consumption patterns (P/Q consumption, On/Off cycle, and Time of Use) of each major appliance were created. The AMI data were coupled to the polynomial load models to achieve the goal of major appliances identification. In this area, the potential use of smart meter data for proper detection of appliances operation at the residential service class needs to be explored in more detail, and expanded to other service classes like commercial and industrial customers. In addition, constant impedance-current-power (ZIP) load models were developed of different home appliances that have recently emerged into the market. The importance of updating the load models was emphasized to properly represent the electrical behavior of the new appliances, and much work still needs to be conducted in load modeling.

References

[1] DoE. (2009). *Guidebook for ARRA Smart Grid Program Metrics and Benefits*. Available: http://www.smartgrid.gov/sites/default/files/pdfs/metrics_guidebook.pdf

[2] RECOVERY.GOV. *Goals of the Recovery Act*. Available: http://www.recovery.gov/About/GetStarted/Pages/WhatisRecoveryAct.aspx

[3] DoE, "Smart Grid Investment Grant Program Funding Opportunity Number: DE-FOA-0000058," U. S. Department of Energy2009.

[4] DoE. *Smart Grid Investment Grant Topic Areas*. Available: http://energy.gov/oe/downloads/smart-grid-investment-grant-topic-areas

[5] DoE. *Recovery Act Selections for Smart Grid Investment Grant Awards - By State - Updated November 2011*. Available: http://energy.gov/oe/downloads/recovery-act-selections-smart-grid-investment-grant-awards-state-updated-november-2011

[6] eMeter, "Understanding the Potential of Smart Grid Data Analytics - A GTM Research Whitepaper."

[7] "Smart Meters and Smart Meter Systems: A Metering Industry Perspective."

[8] FERC. (2012). *Assessment of Demand Response & Advanced Metering Staff Report*. Available: http://www.ferc.gov/legal/staff-reports/12-20-12-demand-response.pdf

[9] NETL, "Advanced Metering Infrastructure Conducted by the National Energy Technology Laboratory for the U.S. Department of Energy Office of Electricity Delivery and Energy Reliability," February 2008.

[10] L. Arthur. What Does It Take to Turn Big Data Into Big Dollars? *Forbes*. Available: http://www.forbes.com/sites/lisaarthur/2012/03/27/what-does-it-take-to-turn-big-data-into-big-dollars/

[11] B. Franks, *Taming The Big Data Tidal Wave: Finding Opportunities in Huge Data Streams with Advanced Analytics*: John Wiley & Sons, Inc., 2012.

[12] M. Schroeck, R. Shockley, J. Smart, D. Romero-Morales, and P. Tufano. Analytics: The real-world use of big data - How innovative enterprises extract value from uncertain data. Available: http://www-03.ibm.com/systems/hu/resources/the_real_word_use_of_big_data.pdf

[13] K. Gering. (2012, July/August 2012) Grid Optimization Built on Smart Meter Networks and Data. *intelligentutility*.

[14] M. Kantardzic, *Data mining : concepts, models, methods, and algorithms*. Hoboken, NJ: Wiley-Interscience : IEEE Press, 2003.

[15] J. A. Jardini, C. M. V. Tahan, M. R. Gouvea, S. U. Ahn, and F. M. Figueiredo, "Daily load profiles for residential, commercial and industrial low voltage consumers," *Power Delivery, IEEE Transactions on,* vol. 15, pp. 375-380, 2000.

[16] J. Han, M. Kamber, J. Pei, and Books24x7 Inc. (2012). *Data mining concepts and techniques, third edition (3rd ed.)* [Text].

[17] T. Dasu and T. Johnson, *Exploratory data mining and data cleaning*. New York: Wiley-Interscience, 2003.

[18] A. Famili, W.-M. Shen, R. Weber, and E. Simoudis, "Data preprocessing and intelligent data analysis," *Intelligent Data Analysis,* vol. 1, pp. 3-23, 1997.

[19] P. J. Huber, *Data analysis : what can be learned from the past 50 years*. Hoboken, N.J.: Wiley, 2011.

[20] J. Grzymala-Busse and W. Grzymala-Busse, "Handling Missing Attribute Values," in *Data Mining and Knowledge Discovery Handbook*, O. Maimon and L. Rokach, Eds., ed: Springer US, 2010, pp. 33-51.

[21] I. Ben-Gal, "Outlier Detection," in *Data Mining and Knowledge Discovery Handbook*, O. Maimon and L. Rokach, Eds., ed: Springer US, 2010, pp. 117-130.

[22] NOAA. *Why Do We have Seasons?* Available: http://www.crh.noaa.gov/lmk/?n=seasons

[23] R. Xu and D. C. Wunsch, *Clustering - IEEE Press Series on Computational Intelligence*. Piscataway, N.J. Hoboken, N.J.: IEEE Press, John Wiley & Sons, Inc., 2009.

[24] B. Everitt, *Cluster analysis*, 5th ed. Chichester, West Sussex, U.K.: Wiley, 2011.

[25] A. K. Jain, "Data clustering: 50 years beyond K-means," *Pattern Recognition Letters,* vol. 31, pp. 651-666, 2010.

[26] G. Gan, C. Ma, and J. Wu, *Data clustering : theory, algorithms, and applications*: SIAM, American Statistical Association, 2007.

[27] A. K. Jain, M. N. Murty, and P. J. Flynn, "Data clustering: a review," *ACM Comput. Surv.,* vol. 31, pp. 264-323, 1999.

[28] B. G. Mirkin, *Clustering for data mining : a data recovery approach*. Boca Raton, FL: Chapman & Hall/CRC, 2005.

[29] S. Das, A. Abraham, and A. Konar, *Metaheuristic clustering*. Berlin: Springer, 2009.

[30] F. Azuaje and J. Dopazo, *Data analysis and visualization in genomics and proteomics*. Chichester, West Sussex ; Hoboken, NJ: John Wiley, 2005.

[31] J. V. d. Oliveira and W. Pedrycz, *Advances in fuzzy clustering and its applications*. Chichester: Wiley, 2007.

[32] B. J. Frey and D. Dueck, "Clustering by Passing Messages Between Data Points," *Science,* vol. 315, pp. 972-976, February 16, 2007 2007.

[33] M. Mézard, "Where Are the Exemplars?," *Science,* vol. 315, pp. 949-951, February 16, 2007 2007.

[34] M. Halkidi, Y. Batistakis, and M. Vazirgiannis, "On Clustering Validation Techniques," *J. Intell. Inf. Syst.,* vol. 17, pp. 107-145, 2001.

[35] J. Handl, J. Knowles, and D. B. Kell, "Computational cluster validation in post-genomic data analysis," *Bioinformatics,* vol. 21, pp. 3201-3212, 2005.

[36] A. K. Jain and R. C. Dubes, *Algorithms for clustering data*. Englewood Cliffs, NJ: Prentice Hall, 1988.

[37] Q. Zhao, "Cluster Validity in Clustering Methods," Ph.D. Dissertation, Faculty of Science and Forestry, School of Computing, University of Eastern Finland, 2012.

[38] F. McLoughlin, A. Duffy, and M. Conlon, "Characterising domestic electricity consumption patterns by dwelling and occupant socio-economic variables: An Irish case study," *Energy and Buildings,* vol. 48, pp. 240-248, 2012.

[39] K. Liu, S. Subbarayan, R. R. Shoults, M. T. Manry, C. Kwan, F. L. Lewis, and J. Naccarino, "Comparison of very short-term load forecasting techniques," *Power Systems, IEEE Transactions on,* vol. 11, pp. 877-882, 1996.

[40] J. W. Taylor, "An evaluation of methods for very short-term load forecasting using minute-by-minute British data," *International Journal of Forecasting,* vol. 24, pp. 645-658, 2008.

[41] G. Che, P. B. Luh, L. D. Michel, W. Yuting, and P. B. Friedland, "Very Short-Term Load Forecasting: Wavelet Neural Networks With Data Pre-Filtering," *Power Systems, IEEE Transactions on,* vol. 28, pp. 30-41, 2013.

[42] M. T. Hagan and S. M. Behr, "The Time Series Approach to Short Term Load Forecasting," *Power Systems, IEEE Transactions on,* vol. 2, pp. 785-791, 1987.

[43] J. W. Taylor, "Short-Term Load Forecasting With Exponentially Weighted Methods," *Power Systems, IEEE Transactions on,* vol. 27, pp. 458-464, 2012.

[44] C. Bo-Juen, C. Ming-Wei, and L. Chih-Jen, "Load forecasting using support vector Machines: a study on EUNITE competition 2001," *Power Systems, IEEE Transactions on,* vol. 19, pp. 1821-1830, 2004.

[45] R. J. Hyndman and F. Shu, "Density Forecasting for Long-Term Peak Electricity Demand," *Power Systems, IEEE Transactions on,* vol. 25, pp. 1142-1153, 2010.

[46] G. Gross and F. D. Galiana, "Short-term load forecasting," *Proceedings of the IEEE,* vol. 75, pp. 1558-1573, 1987.

[47] T. Hong, "Short Term Electric Load Forecasting," PhD dissertation, Operations Research and Electrical Engineering, North Carolina State University, Raleigh, North Carolina, 2010.

[48] "Day-Ahead/Hour-Ahead Forecasting for Demand Trading: A Guidebook," EPRI, Palo Alto, CA, 2001.

[49] H. K. Alfares and M. Nazeeruddin, "Electric load forecasting: Literature survey and classification of methods," *International Journal of Systems Science,* vol. 33, pp. 23-34, 2002/01/01 2002.

[50] W. Charytoniuk and C. Mo-Shing, "Very short-term load forecasting using artificial neural networks," *Power Systems, IEEE Transactions on,* vol. 15, pp. 263-268, 2000.

[51] F. Javed, N. Arshad, F. Wallin, I. Vassileva, and E. Dahlquist, "Forecasting for demand response in smart grids: An analysis on use of anthropologic and structural data and short term multiple loads forecasting," *Applied Energy,* vol. 96, pp. 150-160, 2012.

[52] A. Marinescu, C. Harris, I. Dusparic, S. Clarke, and V. Cahill, "Residential electrical demand forecasting in very small scale: An evaluation of forecasting methods," in

Software Engineering Challenges for the Smart Grid (SE4SG), 2013 2nd International Workshop on, 2013, pp. 25-32.

[53] M. Chaouch, "Clustering-Based Improvement of Nonparametric Functional Time Series Forecasting: Application to Intra-Day Household-Level Load Curves," *Smart Grid, IEEE Transactions on,* vol. PP, pp. 1-9, 2013.

[54] H. S. Hippert, C. E. Pedreira, and R. C. Souza, "Neural networks for short-term load forecasting: a review and evaluation," *Power Systems, IEEE Transactions on,* vol. 16, pp. 44-55, 2001.

[55] S. S. Haykin, *Neural networks and learning machines*, 3rd ed. Upper Saddle River, NJ: Prentice Hall, 2009.

[56] H. Demuth, M. Beale, and M. Hagan, "Neural Network Toolbox (TM) 6 User's Guide - MATLAB," The MathWorks2009.

[57] F. Shu, C. Luonan, and L. Wei-Jen, "Short-Term Load Forecasting Using Comprehensive Combination Based on Multimeteorological Information," *Industry Applications, IEEE Transactions on,* vol. 45, pp. 1460-1466, 2009.

[58] H. S. Hippert and J. W. Taylor, "An evaluation of Bayesian techniques for controlling model complexity and selecting inputs in a neural network for short-term load forecasting," *Neural Networks,* vol. 23, pp. 386-395, 2010.

[59] F. Shu and R. J. Hyndman, "Short-Term Load Forecasting Based on a Semi-Parametric Additive Model," *Power Systems, IEEE Transactions on,* vol. 27, pp. 134-141, 2012.

[60] (August 4, 2009). *Con Edison Launches Smart Grid Pilot Program in Queens*. Available: http://www.coned.com/newsroom/news/pr20090804_2.asp

[61] CER. (November 10, 2013). *Smart Metering Trial Data Publication*. Available: http://www.cer.ie/en/information-centre-reports-and-publications.aspx?article=5dd4bce4-ebd8-475e-b78d-da24e4ff7339

[62] *Data from the Commission for Energy Regulation*. Available: http://www.ucd.ie/issda/data/commissionforenergyregulationcer/

[63] "Effects of Voltage on Load - Draft Report," Dec. 13, 2003.

[64] A. Alkan, "Characterization of Loads Under Varying Voltage Conditions," M.Sc., Electrical Engineering, Polytechnic Institute of New York University, January 2012.

[65] F. L. Quilumba, L. Wei-Jen, H. Heng, D. Y. Wang, and S. Robert Louis, "Load model development for next generation appliances," in *Industry Applications Society Annual Meeting (IAS), 2011 IEEE*, 2011, pp. 1-7.

[66] K. P. Schneider and J. C. Fuller, "Detailed end use load modeling for distribution system analysis," in *Power and Energy Society General Meeting, 2010 IEEE*, 2010, pp. 1-7.

[67] "National Electrical Code - NFPA 70."

[68] *Appliance Efficiency Database*. Available: http://www.appliances.energy.ca.gov/

[69] EnergyStar. *Find ENERGY STAR Products*. Available: http://www.energystar.gov/index.cfm?c=products.pr_find_es_products

[70] *Estimating Appliance and Home Electronic Energy Use*. Available: http://www.energysavers.gov/your_home/appliances/index.cfm/mytopic=10040

[71] C. R. Sastry, V. Srivastava, R. Pratt, and S. Li, "Use of Residential Smart Appliances for Peak-Load Shifting and Spinning Reserves: Cost/Benefit Analysis," Pacific Northwest National Laboratory, December 2010.

[72] R. G. Pratt, C. C. Conner, B. A. Cooke, and E. E. Richman, "Metered end-use consumption and load shapes from the ELCAP residential sample of existing homes in the Pacific Northwest," *Energy and Buildings,* vol. 19, pp. 179-193, 1993.

[73] *Building America - Resources for Energy Efficient Homes: Analysis Spreadsheets*. Available: http://www1.eere.energy.gov/buildings/building_america/analysis_spreadsheets.html

[74] B. Hendron and C. Engebrecht, "Building America House Simulation Protocols," National Renewable Energy Laboratory for the U.S. Department of Energy Building Technologies Program, October 2010.

[75] *GridLAB-D Simulation Software Repository Files - Version 2.2*. Available: http://gridlab-d.svn.sourceforge.net/viewvc/gridlab-d/branch/2.2/residential/

[76] J. Z. Kolter and M. J. Johnson, "REDD: A Public Data Set for Energy Disaggregation Research," *Proceedings of the SustKDD workshop on Data Mining Applications in Sustainability*.

[77] K. Anderson, A. F. Ocneanu, D. Benítez, D. Carlson, A. Rowe, and M. Bergés, "BLUED: A Fully Labeled Public Dataset for Event-Based Non-Intrusive Load Monitoring Research," *Proceedings of the 2nd KDD Workshop on Data Mining Applications in Sustainability*.

[78] T. Peffer, W. Burke, and D. Auslander, "ResPoNSe: modeling the wide variability of residential energy consumption," 2010.

Biographical Information

Franklin L. Quilumba-Gudiño received the Diploma degree in Electrical Engineering from the National Polytechnic School (Escuela Politécnica Nacional, EPN), Quito-Ecuador, in 2008. He held a teaching position at the EPN from 2008 to 2009. He is currently pursuing the PhD degree in the area of steady-state and dynamic analysis of power systems at the University of Texas at Arlington (UTA). He is a member of the research group at the Energy Systems Research Center at UTA. His areas of interest are power systems analysis, operation, stability and control; power plants; computer simulation of electric power systems; power load modeling; demand response; and load forecasting.