Mining Hidden Knowledge from Measured Data for Improving Building Energy Performance

Zhun Yu

A Thesis

in

The Department

of

Building, Civil and Environmental Engineering

Presented in Partial Fulfillment of the Requirements
for the Degree of Doctor of Philosophy at
Concordia University
Montréal, Québec, Canada

January, 2012

© Zhun Yu, 2012

CONCORDIA UNIVERSITY
SCHOOL OF GRADUATE STUDIES

This is to certify that the Ph.D. thesis prepared

By: **Zhun Yu**

Entitled: **Mining Hidden Knowledge from Measured Data for Improving Building Energy Performance**

and submitted in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

complies with the regulations of this University and meets the accepted standards with respect to originality and quality.

Signed by the final examining committee:

_____ Dr. G. Gouw (Chair)

_____ Dr. C. Inard (External Examiner)

_____ Dr. W. Ghaly (External to Program)

_____ Dr. R. Zmeureanu (Examiner)

_____ Dr. Z. Chen (Examiner)

_____ Dr. B. Fung (Thesis Co-supervisor)

_____ Dr. F. Haghighat (Thesis Supervisor)

Approved by _____
Chair of Department or Graduate Program Director

_____2012_____
Dean, Faculty of Engineering and Computer Science

ABSTRACT

Mining Hidden Knowledge from Measured Data for Improving Building Energy Performance

Zhun Yu, Ph.D.
Concordia University, 2012

Nowadays, building automation and energy management systems provide an opportunity to collect vast amounts of building-related data (e.g., climatic data, building operational data, etc.). The data can provide abundant useful knowledge about the interactions between building energy consumption and its influencing factors. Such interactions play a crucial role in developing and implementing control strategies to improve building energy performance. However, the data is rarely analyzed and this useful knowledge is seldom extracted due to a lack of effective data analysis techniques.

In this research, data mining (classification analysis, cluster analysis, and association rule mining) is proposed to extract hidden useful knowledge from building-related data. Moreover, a data analysis process and a data mining framework are proposed, enabling building-related data to be analyzed more efficiently. The applications of the process and framework to two sets of collected data demonstrate their applicability. Based on the process and framework, four data analysis methodologies were developed and applied to the collected data.

Classification analysis was applied to develop a methodology for establishing building energy demand predictive models. To demonstrate its applicability, the

methodology was applied to estimate residential building energy performance indexes by modeling building energy use intensity (EUI) levels (either high or low). The results demonstrate that the methodology can classify and predict the building energy demand levels with an accuracy of 93% for training data and 92% for test data, and identify and rank significant factors of building EUI automatically.

Cluster analysis was used to develop a methodology for examining the influences of occupant behavior on building energy consumption. The results show that the methodology facilitates the evaluation of building energy-saving potential by improving the behavior of building occupants, and provides multifaceted insights into building energy end-use patterns associated with the occupant behavior.

Association rule mining was employed to develop a methodology for examining all associations and correlations between building operational data, thereby discovering useful knowledge about energy conservation. The results show there are possibilities for saving energy by modifying the operation of mechanical ventilation systems and by repairing equipment.

Cluster analysis, classification analysis, and association rule mining were combined to formulate a methodology for identifying and improving occupant behavior in buildings. The results show that the methodology was able to identify the behavior which needs to be modified, and provide occupants with feasible recommendations so that they can make required decisions to modify their behavior.

ACKNOWLEDGEMENTS

First of all, I would like to express my deepest gratitude to my supervisor, Dr. F. Haghighat, for his excellent guidance throughout my graduate studies. I benefit greatly from both his extensive expertise and charismatic personality. His strong support and great patience enabled me to strive for the highest level of achievement in carrying out this work, which led to a number of research accomplishments.

I would also like to extend my sincere gratitude and appreciation to my co-supervisor, Dr. B.C.M. Fung, for his valuable suggestions on my research and tremendous efforts to improve the quality of this thesis. Without his help, I could not finish this thesis.

Additionally, I am deeply indebted to all other supervisory committee members, Dr. G. Gouw, Dr. C. Inard, Dr. R. Zmeureanu, Dr. W.S. Ghaly, and Dr. Z. Chen, for accepting the appointment to the dissertation committee, as well as for their suggestions and support.

Financial support by Public Works and Government Service Canada (PWGSC), and Concordia University is gratefully acknowledged. I wish to express my gratitude to Dr. E.Morosfsky from PWGSC for having the vision to support my studies.

Many thanks to the School of Graduate Studies of Concordia University for providing me with valuable scholarships, awards and assistantships during my Ph.D. studies.

Special acknowledgements go to Dr. Yoshino Hiroshi, Yves Gilbert, and Denis Dumont for providing building-related data.

Special thanks to Dr. Liang (Grace) Zhou for her support, encouragement, and help

with English writing.

I also thank my colleagues: Dr. Jiang Zhang, Dr. Parham Mirzaee, Lexuan Zhong, Arash Soleimani, Omid Ashrafi, Arash Bastani, Vida Safari, Reza Mostofi. They created a good and relaxed atmosphere in our office so that I can focus on and enjoy my research.

I would like to thank all my friends in Montreal, for their help, friendship and encouragements.

Finally, my utmost gratitude goes to my parents and brother, for their endless love and support without which this work would have been impossible.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

NOMENCLATURE

End Use Loads

HVAC	Heating, ventilation, air-conditioning load	(MJ/m^2)
SHW	Supply hot water load	(MJ/m^2)
KITC	Kitchen load	(MJ/m^2)
LIGHT	Lighting load	(MJ/m^2)
REF	Refrigerator load	(MJ/m^2)
A&I	Amusement and information load	(MJ/m^2)
H&S	Housework and sanitary load	(MJ/m^2)
OTHER	Other load	(MJ/m^2)

English letters

T	Annual mean air temperature	(°C)
RH	Annual mean relative humidity	
V	Annual mean wind speed	(m/s)
RA	Annual mean global solar radiation	(MJ/m^2)
HT	House types	
CO	Construction type	
FA	Building area	(m^2)
ELA	Equivalent leakage area	(cm^2/m^2)
HLC	Heat loss coefficient	(W/m^3K)

NO	Number of occupants	
HEAT	Space heating and cooling	
HWS	Hot water supply	
KITC	Kitchen equipment	
TA	Air temperature	(°C)
TG	Glycol temperature	(°C)
H	Relative humidity	(kg/kg)
Q	Flow rate	(L/S)
F	Frequency of variable-speed drives on fans	(Hz)

Subscripts

I, II, III, IV, V	Fresh air handling unit 1 (FHU1), FHU2, FHU3, FHU4, FHU5
VI, VII	Return air handling unit 1 (RHU1), RHU2
VIII, IX	Exhaust air handling unit 1 (EHU1), EHU2
ac	After cooling coil
ah	After heating coil
br	Before recuperation
ar	After recuperation
1, 2, 3	Fan1, fan2, fan3
i, ii, iii	Recuperation1, recuperation2, recuperation3
o	Outdoor

VA	VA part
ENCS	ENCS part

1. INTRODUCTION

1.1 Background and Motivation

Energy consumed in the building sector is of growing concern. With rising living standards, building energy consumption has significantly increased over the past few decades. For example, from 1994 to 2004, building energy consumption in Europe and North America increased at a rate of 1.5% and 1.9% per annum, respectively (Hein, 2005; Pérez-Lombard et al., 2008). Building energy consumption in China has increased more than 10% per annum for the past 20 years (Cai, 2009). The high and steady increase in demand for energy necessitate a thorough understanding of the major influencing factors to assist in developing effective approaches to reducing building energy consumption. Factors influencing building energy consumption can be divided into seven categories (Yu et al., 2011):

(1) Climate (e.g., outdoor air temperature, solar radiation, wind velocity, etc.);

(2) Building-related characteristics (e.g., type, area, orientation, etc.);

(3) User-related characteristics, except for social and economic factors (e.g., user presence, etc.);

(4) Building services systems and operation (e.g., space cooling/heating, hot water supplying, etc.);

(5) Building occupants' behavior and activities;

(6) Social and economic factors (e.g., degree of education, energy cost, etc.); and

(7) Indoor environmental quality required.

These seven factors play an essential role in reducing energy consumption and should be clearly understood. However, there still are significant barriers that prevent researchers and architects from achieving the goal of completely understanding these factors. For example, researchers and architects often observe a large discrepancy between the designed/simulated and actual building energy consumption, and they are unable to give a clear explanation for this discrepancy. Another challenge is to clearly identify the effects of these influencing factors, especially occupant behavior, on building energy consumption. These barriers can lead to misunderstandings of how the influencing factors will affect building energy performance, and thus add difficulties to energy consumption reduction. Therefore, it is vital that these barriers are removed so that building energy performance can be improved efficiently.

To overcome these barriers, one effective method is to analyze measured building-related data and acquire relevant useful knowledge, considering that such data contains actual knowledge about these influencing factors. In general, building-related data includes (Yu et al., 2011):

(1) Climatic data (e.g., outdoor air temperature, outdoor relative humidity, etc.);

(2) Building operational data, mainly operational data of HVAC systems (e.g., supply air temperature, fresh air flow rates, etc.), IEQ data (e.g., indoor air temperature, human

thermal comfort, etc.), and energy data (e.g., monthly electricity consumption, end-use loads of household appliances, etc.); and

(3) Building physical parameters (e.g., floor area, window-to-wall-ratio, etc.);

Currently, vast amounts of building-related data have been collected and stored, since building automation systems (BAS) are extensively employed. Moreover, for an existing building, building-related data can be surveyed through different methods (e.g., analysis of design documentation, questionnaires, and interviews). This data contains abundant knowledge of building design, operation, and maintenance that can be extracted to help reduce building energy consumption. However, the data is rarely analyzed and translated into useful knowledge, mainly due to its complexity (especially large volumes and poor quality) and a lack of effective data analysis techniques. Consequently, this motivated this study with the purpose of establishing a data analysis process and a systematic data analysis framework, to deal with the challenges caused by the complexity of measured building-related data. Note that the data analysis process refers to a series of sequential steps in analyzing measured building-related data. The data analysis framework mainly includes different data analysis algorithms, from which a set of efficient data analysis methodologies can be developed. Both the process and the framework are aimed at successfully mining hidden and useful knowledge from measured building-related data in order to improve building energy performance.

1.2 Problem Statement

Various data analysis techniques, especially traditional statistical analysis and building simulation, have been widely used in building-related studies. One main goal is to analyze the complex interactions between building energy consumption and its influencing factors, thereby improving building energy performance. However, considering the increased size of building historical databases and the diversity of the influencing factors, these commonly-used data analysis methods are insufficient to take full advantage of measured building-related data to account for the interactions and help improve performance. In particular, a number of problems of building energy performance improvement remain significant barriers to researchers and architects; and these problems are difficult to completely solve by using these commonly-used data analysis methods. Four fundamental problems can be listed as follows:

(1) How can we develop reliable building energy–demand models that are interpretable and that can be easily used by people without advanced mathematical knowledge?

(2) How can we investigate building occupant behavior and quantitatively identify its effect on building energy consumption without including the impact of other influencing factors such as weather conditions?

(3) How can we examine all the associations and correlations among building operational data (e.g., various operational parameters of HVAC systems), and acquire useful knowledge from them to better understand building operation and

reduce energy consumption?

(4) How can we identify energy-related occupant behavior that needs to be modified for energy conservation, and how can we make recommendations for behavior modification?

Clearly, in order to take advantage of measured building-related data and address these problems, it is necessary to propose more effective data analysis techniques and extract relevant useful knowledge from the data. Furthermore, it is highly desirable to provide an avenue for standardizing the process of data analysis within the building engineering domain. Researchers and architects will greatly benefit from a standardized process that enables them to efficiently analyze measured building-related data and obtain useful knowledge about improving building energy performance. Accordingly, a data analysis process and a systematic data analysis framework need to be established based on the proposed data analysis techniques.

1.3 Proposed Data Analysis Techniques

In this research, data mining is proposed as a primary tool to analyze measured building-related data. Data mining techniques excel at automatically analyzing huge amounts of data for useful information and fit well with the purpose of this research.

In the past decade, different definitions of data mining have been given by various researchers. For example, Hand et al. (2001) define data mining as "the analysis of large observational data sets to find unsuspected relationships and to summarize the data in

novel ways so that data owners can fully understand and make use of the data." As defined by Cabena et al. (1998), data mining is "an interdisciplinary field bringing together techniques from machine learning, pattern recognition, statistics, databases, and visualization to address the issue of information extraction from large databases." Based on these statements, it can be concluded that data mining is essentially a combination of multi-disciplinary approaches. It is often used to extract hidden but useful patterns from a large volume of data and to transform the data into knowledge that could benefit further work. Data mining has been successfully applied in many scientific, medical, and application domains (e.g., banking, bioinformatics, new materials identification, fraud detection, and telecommunications). It was also identified by the *MIT Technology Review* (MIT Technology Review, 2001) as one of the ten emerging technologies that may change the world. In this study, three widely accepted and implemented data mining techniques were employed: data classification, clustering analysis, and association rule mining. Each of these techniques will be further discussed in the following chapters.

1.4 Purpose and Objectives

The purpose of this research is to construct a data analysis process and a systematic data mining framework within the building engineering domain and then validate them. The proposed process and framework can help to analyze measured building-related data and discover useful knowledge for evaluating and improving building energy performance. The process describes knowledge extraction from measured data step by step. As one

major step in the process, the framework helps to develop different data analysis methodologies based on selected data mining techniques and algorithms. These methodologies can be applied to deal with various ranges of problems within the building engineering domain. To demonstrate the applicability of the proposed process and framework, measured data from selected buildings is collected and analyzed; a set of data analysis methodologies are developed to address the four fundamental problems outlined in Section 1.2.

The main objectives of this thesis are:

(1) To develop a data analysis methodology that establishes reliable building energy-demand models that are interpretable and that can be easily used by people without training in advanced mathematics and statistics;

(2) To develop data analysis methodologies for studying building occupant behavior, such as quantitatively identifying the effect of occupant behavior on building energy consumption, and identifying the occupant behavior that can be modified to save energy; and

(3) To develop a data analysis methodology that examines all the associations and correlations among building operational data, and extracts useful knowledge from them to better understand system operation and reduce energy consumption.

1.5 Organization of the Thesis

This chapter has introduced the research background and motivation, the problem

statements, the proposed data analysis techniques, and the purpose/objectives of this research.

Chapter 2 reviews existing data analysis methods for extracting useful knowledge from measured building-related data and the application of three data mining techniques within the domain of building engineering. Also, the literature review of the data analysis methods for addressing the four abovementioned problems is conducted.

Chapter 3 introduces the proposed data analysis process and framework, as well as the three data mining techniques. Then, collected data for the case studies in this research is described.

In Chapter 4, the application of a basic data mining technique (i.e., data classification) to establish building energy-demand models is presented and discussed.

In Chapter 5, the development of a methodology for examining the influences of occupant behavior on building energy consumption is reported. The method is based on a basic data mining technique: cluster analysis.

In Chapter 6, a methodology for examining all the associations and correlations between building operational data is proposed for discovering useful knowledge about energy conservation. The method is based on a basic data mining technique (association rule mining).

In Chapter 7, a methodology is developed for identifying the occupant behavior that needs to be modified in existing residential buildings. The method is based on the three

data mining techniques: data classification, cluster analysis, and association rule mining.

Chapter 8 concludes the thesis and proposes future work.

2. LITERATURE REVIEW

This chapter evaluates the data analysis methods for extracting useful knowledge from building-related data and reviews the applications in the domain of building engineering of the three main data mining techniques: data classification, clustering analysis, and association rule mining. The methods used by previous researchers who attempted to solve the four problems outlined in Chapter 1 are then summarized and assessed.

2.1 Data Analysis Methods for Improving Building Energy Performance

MacDonald and Wasserman (1989) summarized five general categories of data analysis methods employed for evaluating and improving building energy performance based on measured building-related data as follows:

(1) *Annual total energy use* and *energy use intensity* (EUI) comparison,

(2) Linear regression and component models,

(3) Multiple linear regression models,

(4) Building simulation programs (also termed microdynamic modeling), and

(5) Dynamic thermal performance models (also termed macrodynamic modeling).

In the first analysis method, both *annual total energy use* and the EUI are typical energy performance indicators. Accordingly, the method can be categorized as the typical

indicators method, given that other indicators (e.g., *coefficient of performance*) may also be used.

The second and third methods relate to regression analysis, a statistical technique. Hence, they are merged into the same category for simplicity and categorized as the statistical analysis method, given that other statistical techniques such as correlation analysis) may also be used. Similarly, the fourth and fifth methods can be merged and categorized as the building simulation method.

Consequently, the five categories are merged into three to better describe the data analysis methods for extracting useful knowledge from measured building-related data:

(1) Typical indicators,

(2) Statistical analysis, and

(3) Building simulation.

In the following section, each method is reviewed and evaluated.

2.1.1 Typical Indicators Method

Typical indicators, such as *annual total energy use* and the EUI, are a simple method of analyzing measured building-related data and evaluating building energy performance. *Annual total energy use* refers to the building energy consumption in one year. The EUI is a measure of energy efficiency and is calculated as the ratio of *annual total energy use* to an total floor area. These two indicators were mainly utilized to survey building energy-use

patterns and investigate the impact of the influencing factors of building energy consumption (Deng and Burnett, 2000; AboulNaga and Elsheshtawy, 2001; Deng 2003; Balaras et al., 2007; Chen et al., 2009; Filippín et al., 2009; Chung and Hui, 2009; Priyadarsini et al., 2009). Also, these indicators could be utilized to compare the building energy consumption before and after retrofitting, thereby evaluating the energy-saving potential of various energy-saving techniques and energy efficiency improvements (Santamouris et al., 1996; Balaras et al., 2002; Balaras et al., 2003). Other similar indicators, such as *annual total heating/cooling energy consumption* and *annual total energy supply cost*, were also utilized for data analysis. For example, Long and Zhou (2005) studied the influence of shading measures on building energy consumption using both *annual heating energy consumption* and *annual cooling energy consumption*. Li et al. (2006) designed a distributed combined heating, cooling, and power generation system in Beijing, with thermal performance, economics, and environment factors being considered simultaneously based on *annual total energy supply cost*.

The major advantage of the typical indicators method is its simplicity. Moreover, the use of these typical indicators makes it possible to compare different designs. However, typical indicators alone are insufficient to analyze measured building-related data and evaluate building energy performance. Particularly, they cannot provide insights into building energy-use patterns.

2.1.2 Statistical Analysis Method

Statistical analysis techniques, particularly regression analysis (including both linear regression and non-linear regression), were extensively used within the building engineering domain. Regression analysis was utilized to identify the correlation between building energy consumption and its influencing factors (e.g., climate, occupancy patterns, HVAC system design and operation, and building physical parameters), and then to analyze overall building energy-use patterns and how these influencing factors affect energy consumption (Hammarsten, 1979; Monts, and Blissett, 1982; Gaunt, 1985; Zmeureanu and Fazio, 1991; Deng and Burnett, 2000; Yu and Chow, 2001; Deng 2003; Tonooka et al., 2006; De la Flor et al., 2006; Chung and Hui, 2009; Priyadarsini et al., 2009; Chen et al., 2009). An additional application of regression analysis was to predict building energy demand based on environmental data or building physical parameters (Sullivan and Nozaki, 1984; Sullivan et al., 1985; O'Neill et al., 1991; Lam et al., 1997; Dong et al., 2005; Chung and Hui, 2009). Also, regression analysis was used to predict other parameters, such as indoor air temperature and relative humidity (Givoni and Krüger, 2003; Krüger and Givoni, 2004; Freire et al., 2008), the overall heat transfer coefficient (the U-value) (Jiménez and Heras, 2005), and the energy consumption of different types of cooling plants (e.g., centrifugal chillers and ice storage systems (Kim and Kim 2007)). Additionally, some researchers compared building energy performance in different countries or cities by using statistical techniques. For example, Zhang (2004) compared

residential energy-use patterns in China with those in Japan, Canada, and the United States by using relationships between energy consumption and heating degree-days.

The strength of statistical techniques is their simplicity and widespread familiarity. However, most statistical techniques are utilized with the premise that data analysts, based on their expertise, "believe" that strong associations and correlations exist among two or more parameters. For example, researchers perform regression/correlation analysis between building energy consumption and outdoor air temperature because they "believe" that outdoor air temperature may have a significant influence on building energy consumption. Such analysis depends mainly on the prior expertise of analysts and adopted statistical techniques. As a result, useful knowledge could be lost, particularly indirect associations and correlations between data (e.g., parameters A and B do not have a direct impact on C, but they may have an indirect impact through parameters D and E) (Yu et al., 2011).

2.1.3 Building Simulation Method

Building simulation is another method widely employed to analyze measured building-related data. Various simulation programs, such as EnergyPlus (Crawley et al., 2001) and TRNSYS (Al-ajmi and Hanby, 2008), were commonly utilized when using this method. In some cases this method was used to conduct building energy consumption calculations in order to identify the correlations between building energy consumption and

different influencing factors (e.g., total building energy consumption and building relative compactness (Ourghi et al., 2007), heating/cooling loads and building control strategies (Eskin and Türkmen, 2008), and annual electricity consumption and the overall heat-transfer coefficient U (Lam, 2000)). In other cases, the energy-saving potential of various energy conservation techniques, such as green building design options (Pan et al., 2008), building-integrated photovoltaic (PV) technologies (Ordenes et al., 2007), and PV ventilated window systems (Chow et al, 2007), were evaluated using this method. Additionally, some researchers used simulation programs to model the energy consumption of various building services systems, and then compared the actual energy consumption with simulated results to evaluate the performance of those systems (Lazzarin et al., 2005; Zhou et al., 2008; Tian and Love, 2009; Li et al., 2010).

Building simulation allows for the prediction of building energy performance under various conditions. However, this method does not perform well in simulating energy performance for occupied buildings as compared to non-occupied buildings, due to a lack of sufficient knowledge about occupant behavior patterns, which are normally very complicated. Additionally, the application of building simulation programs is normally complicated and the learning process is time-consuming (Yu et al., 2011).

2.2 Application of the Three Data Mining Techniques in Building Engineering

In this study, three data mining techniques—data classification, cluster analysis, and

association rule mining—are proposed as primary methods for mining hidden and useful knowledge from measured building-related data. These techniques have been extensively applied in various fields such as industrial and medical studies (Delgado et al., 2001; Jiao and Zhang, 2005; Georgilakis et al., 2007; Pan et al., 2007; Hsu, 2009). However, their utilization within the domain of building engineering is still sparse. It should be mentioned that, due to the fact that several classification methods (e.g., ANN method, Genetic Algorithm, Rough Set approach, and Fuzzy Set approach) were less commonly used for data classification in commercial data mining systems, in this research these methods were not assigned to data classification (but were still included in the data mining system).

In particular, previous work seldom studied how to utilize these three data mining techniques to process building-related data and extract useful knowledge. With regard to the association rule mining technique, no literature was found, to the best of our knowledge. With regard to the data classification technique, Tso and Yau (2007) compared the accuracy of regression analysis, the ANN method, and the decision tree method (i.e., one typical data classification method) in predicting the average weekly electricity consumption for both summer and winter in Hong Kong. With regard to the cluster analysis technique, Santamouris et al. (2007) applied the technique to classify and rate the energy performance of school buildings. Based on the cluster analysis and Principal Component Analysis (PCA) techniques, Gaitani et al. (2010) proposed an

approach to rating the energy performance for space heating and evaluating potential energy savings in the school sector in Greece. Also, Lam et al. (2009) combined the cluster analysis and the PCA to identify climatic influences on chiller plant electricity consumption. Wu and Clements-Croome (2007) applied the cluster analysis technique to analyze indoor environmental data measured from wireless sensor networks which was heavily noisy. In their study, cluster analysis was used first to discover outliers and then to estimate the distribution of indoor temperature.

In summary, data mining, especially the three data mining techniques, is relatively a new concept/tool applicable to the building engineering domain. Hence, our study may bring a new inspiration for architects and researchers to find approaches to reducing building energy consumption and realizing the goal of ultra-low energy consumption in buildings. Furthermore, if a process and a systematic framework of data mining application can be established, they will greatly benefit building practice and future analysis.

2.3 Building Energy Prediction Models

In recent years, different models have been developed to predict building energy demand. Generally, these models can be divided into two main categories: regression models and ANN models.

Regression models

Regression models correlate building energy demand with relevant variables such as climatic variables (e.g., outdoor/indoor temperature and relative humidity) and physical

variables (e.g., wall type, building geometry, and window-to-wall-ratio) (Sullivan and Nozaki, 1984; Sullivan and Nozaki, 1985; O'Neill et al., 1991; Lam et al., 1997; Westergren et al., 1999; Dong et al., 2005; Caldera et al., 2008; Chung and Hui, 2009). For example, Catalina et al. (2008) used regression models to predict the monthly heating demand for single-family residential buildings in temperate climates (16 major cities in France). Ghiaus (2006) used a regression technique to study whether the heating losses and the outdoor temperature have the same distribution, and then developed a regression model for predicting the heat losses.

The main advantage of regression models lies in their computational simplicity. However, this method has a severe limitation: building operational data (e.g., operational data of HVAC systems) is usually recorded at short time intervals, which can be considered instantaneous. As a result, various random disturbances that do not usually follow a normal (Gaussian) distribution, such as occupancy, ventilation rates, and solar gains, can add bias and noise to the data, reducing the correlation and prediction accuracy (Ghiaus, 2006).

ANN models

Previous studies showed that ANN models have also been widely applied to correlate building energy demand with climatic/physical variables (Kreider and Wang, 1992; Anstett and Kreider, 1993; Kawashima, 1994; Stevenson, 1994; Kreider et al., 1995; Andersson et al., 1996; Kreider and Wang, 1997; Aydinalp et al., 2002; Yang et al., 2005; Dong et al., 2005; Ekici and Aksoy, 2009; Li et al., 2009). For example, Olofsson et al. (2001)

investigated the potential of a neural network to predict the annual space heating demand of a building, based on the measured average daily outdoor and indoor temperatures and space heating demand for a limited time period. Also, PCA was applied to the measured data for choosing model parameters. The results showed that an ANN was able to produce good predictions except for certain periods when the space heating demand was very low.

The most important advantage of ANN models, over other models, is the ability to provide predictions even for a multivariable mixed-integer problem, which involves both integers (e.g., binary values) and continuous variables (Yao et al., 2006). However, the major limitation of this method is that the network is considered a black-box model—a relationship between the individual influencing factor and output cannot be observed directly.

In summary, a review of the two main energy demand modeling methods was conducted. These two modeling methods have been successfully applied to predict building energy demand. However, regression models involve complicated equations and ANN models operate like a "black box"; therefore, the models developed using these methods are not understandable and interpretable especially for common users without advanced mathematical knowledge. This makes it difficult for these methods to be used as common predictive tools. In order to overcome such limitations, decision tree–based predictive models based on a typical data mining technique (i.e., data classification) are proposed. Generally, the establishment of decision tree–based models does not consider

the correlation among input parameters. Also, decision tree–based models use an interpretable tree structure to provide insights into the relationship between various influencing factors and output. The decision tree method will be introduced in more detail in Chapter 3.

2.4 The Effects of Occupant Behavior on Building Energy Consumption

The identification of major determinants of building energy consumption, together with a thorough understanding of the impacts of the identified determinants on energy consumption patterns, could assist in achieving the goal of improving building energy performance and reducing greenhouse gas emissions due to the building energy consumption. As mentioned previously, factors influencing the total building energy consumption can be divided into seven categories:

(1) Climate. (2) Building-related characteristics. (3) User-related characteristics, except for social and economic factors. (4) Building services systems and operation. (5) Building occupants' behavior and activities. (6) Social and economic factors. (7) Indoor environmental quality required.

Among these seven factors, social and economic factors will partly determine occupants' attitudes toward energy consumption, and building occupants will embody such impact in their daily activities and behavior, thereby influencing building energy consumption. At the same time, indoor environment quality could be regarded as being basically decided by the occupants, thereby influencing building energy consumption. In

essence, these two categories of factors which represent occupants' influences affect building energy consumption indirectly. Therefore, their influences on building energy consumption are already contained within the effects of occupant behavior, and there is no need to take them into consideration when identifying the effects of influencing factors.

The separate and combined influences of the first four factors on building energy consumption can be identified via simulation. With a variety of parameter settings, current simulation software is robust in respect to simulating different situations based upon these four factors. However, it is difficult to completely identify the influences of occupant behavior and activities through simulation due to users' behavior diversity and complexity; current simulation tools can only imitate behavior patterns in a rigid way. In recent years several models have been established to integrate the influence of building occupant behavior into building simulation programs (Reinhart, 2004; Bourgeois, 2005; Rijal et al., 2007; Hoes et al., 2009). However, these models focus only on typical activities such as the control of sun-shading devices, while realistic building user-behavior patterns are more complicated.

A number of studies (Nakagami, 1996; Lopes et al., 2005; Yu et al., 2010) suggest that, to estimate the effects of user behavior, one possible approach is to extract corresponding useful information from measured building-related data. Generally, the previous studies on the effects of occupant behavior can be divided into two categories. The first category focuses on the effects of building user presence on building energy consumption. For

example, Emery and Kippenhan (2006) reported a survey on the effects of occupant presence on home energy usage in four nearly identical houses. The four houses were divided into two pairs, and the building envelope of one pair was constructed with improved thermal resistance. One of each pair of houses was left unoccupied, while the other was occupied. Researchers compared the first heating season's (1987–88) total energy consumption of the occupied and unoccupied houses (i.e., the sum of heating, lighting, and appliances). They found that the presence of occupants increased the total energy consumption of both occupied houses, and the house with the improved building envelope had a smaller increase.

The second category of studies focuses on the effects of actions occupants took to influence energy consumption. For example, Ouyang and Hokao (2009) investigated energy-saving potential by improving user behavior in 124 households in China. These houses were divided into two groups: one group was educated to promote energy-conscious behavior and put corresponding energy-saving measures into effect in July 2008, while the other group was not informed. Comparisons were made between monthly household electricity uses in July 2007 and July 2008 for both groups. Researchers found that the effective promotion of energy-conscious behavior could reduce household electricity consumption by more than 10%.

Evidently, comparative analyses on measured data were conducted in these studies to identify the effects of user behavior. However, the limitations of this method are

significant.

First, apart from user behavior, the other four influencing factors also simultaneously contribute to the variation in building energy consumption, while this method is unable to adequately remove the effects of those four factors and identify the influences of occupant behavior. Although in these studies some measures were implemented to remove the impact of those factors, such as by using nearly identical housing characteristics and by taking energy data in other years with similar climatic conditions as a reference, the effects of these measures are questionable since even a slight difference in some building parameters (e.g., heat loss coefficient) and weather parameters (e.g., annual average outdoor air temperature) would result in remarkable fluctuations in the building energy consumption.

Second, in building databases, buildings are usually described by a mixture of variable types such as numerical variable, categorical variable (e.g., residential building types divided into detached and apartment), and ordinal variable (e.g., buildings rated as platinum, gold, or silver). Such data of mixed variable types is difficult to process by statistical methods that are normally utilized in comparative analyses. This also adds the difficulty of distinguishing between building-related effects and user-related effects.

Third, with regard to comparative analyses, buildings are usually classified into different groups to simplify research. Such classification is commonly based on building-related parameters such as floor area. For example, if building floor area ranges

from 100 m^2 to 400 m^2, it can be classified as small, medium, and large corresponding to the intervals [100, 200), [200, 300), and [300, 400], respectively. The partitioning of building-related parameters is normally decided by considerations of convenience and intuition. Why should 200 m^2 and 300 m^2 be the interval between each group? Hence, a more rational classification method is required for grouping buildings.

Moreover, buildings are commonly represented by various typical parameters at the same time, such as building age and floor area. These parameters may be divided into different levels for simplicity, such as low and high. In order to perform a comprehensive investigation, the sample size (i.e., number of buildings) necessary for research should be determined by the combination of different levels of all parameters. For example, suppose seven typical parameters are selected for representation and each are stratified into 3 levels (e.g., small, medium, and large). Combinatorial theory shows that at least $3^7 = 2187$ buildings should be investigated for comparison, which may be impractical.

In summary, it is difficult to identify the effects of occupant behavior on building energy consumption, since the influence of other energy use determinants cannot be removed. In this research, we propose one of the typical data mining techniques, clustering analysis, for examining the individual effects of occupant behavior. Clustering is an unsupervised learning algorithm. Its goal is to identify a set of previously undefined clusters among data by using special mathematical techniques based on the similarity of the data features. This technique will be introduced in more detail in Chapter 3.

2.5 Discovering Associations and Correlations among Measured Data

Building-related data may have a direct/indirect influence on each other, considering that they are closely related to the same buildings. Specifically, there may be strong associations (i.e., connections or relationships) and correlations between them. Both these associations and correlations should be examined to understand building operation, determine rules of conserving energy, and develop appropriate strategies to design buildings.

A number of studies have been conducted to identify associations and correlations between measured building-related data. Researchers utilized statistical analysis techniques, particularly regression analysis, and focused mainly on the relationships between building energy consumption and its influencing factors, such as building physical parameters (Yu and Chow, 2001; Deng, 2003; Yu et al., 2010), occupancy patterns (Priyadarsini et al., 2009; Yu et al., 2011), building operation and management (Chung and Hui, 2009), social and economic factors (Tonooka et al., 2006), indoor air quality requirements (Chen et al., 2010), and weather conditions (De la Flor et al., 2006). However, few researchers examined associations and correlations between building operational data, especially operational data of HVAC systems, to better understand building operation in order to improve building performance. This is mainly due to the complexity of such data and a lack of effective data analysis techniques. Note that the energy consumption of HVAC systems can account for a large portion of total building

energy consumption (Pérez-Lombard et al., 2011).

The main data analysis methods used to discover associations and correlations among measured data within the building engineering domain (i.e., statistical analysis techniques) were reviewed in Section 2.1.2. The limitation of these techniques was also addressed. Moreover, many parameters are usually monitored and huge amounts of operational data are collected on HVAC systems. Consequently, it is difficult, and often infeasible, for data analysts to conduct statistical analyses, correlation analyses for example, on every combination of parameters in order to discover all of the associations and correlations that are crucial for achieving optimum building performance.

In this research, we propose one of the most widely applied techniques in data mining (i.e., association rule mining) for discovering all the useful and important associations and correlations between building operational data. This technique will be introduced in more detail in Chapter 3.

2.6 Approaches to Modifying Occupant Behavior in Residential Buildings

Recently there has been mounting interest in studying occupant behavior in buildings and in developing methodologies for identifying the corresponding energy-saving potential. As reviewed in Section 2.4, Ouyang and Hokao (2009) investigated the energy-saving potential by improving user behavior in 124 households in China. Al-Mumin et al. (2003) simulated occupant behavior improvement (i.e., occupant behavior before and after

modification) and the corresponding annual electricity consumption reduction by using the energy simulation program ENERWIN. They first collected data and information on occupancy patterns and operation schedules of electrical appliances in 30 selected residences in Kuwait. This data and information were then used in ENERWIN to replace the default value. A house was then selected as a case study, and the simulation results showed that the annual electricity consumption in this house was increased by 21%. The results also indicated that the ENERWIN's default parameters (i.e., parameters taken from the software manual) are probably more appropriate for the Western lifestyle. Moreover, it was found that a 39% reduction in energy consumption can be achieved by improving occupant behavior such as turning off lights when rooms are empty and setting the air conditioner thermostat to a higher temperature (but still within the comfort level).

Two approaches (i.e., energy-saving education and building simulation) were used to modify occupant behavior in residential buildings and identify the corresponding energy-saving potential. These approaches can help to modify occupant behavior and have an immediate effect on the building energy consumption reduction. However, both approaches have certain limitations (Yu et al., 2011).

Regarding the energy-saving education approach, commonly detailed energy-saving measures and tips on the efficient use of various household appliances should be provided for occupants. Considering that a family normally has a number of appliances, and that each appliance may have various tips (e.g., for refrigerators: reduce the number of door

open times and its duration, keep coils and filters clean, position it away from heat sources, etc.), there could be a large number of energy-saving measures and tips for an individual family. For example, one family may have 30 household appliances, with each appliance having an average of 8 energy-saving tips. Accordingly, the occupants need to follow and implement 240 tips, which is impractical. Although a booklet of these tips can be prepared for building occupants, it is very difficult for occupants to remember them all distinctly and implement them for a long time in practice. Furthermore, occupants may not fully understand and have confidence in these tips' effectiveness because they only provide qualitative information. In addition, some energy-saving opportunities can only be initiated by building occupants. For example, when occupants realize they have consumed too much energy on both computers and TVs, they can avoid using both devices simultaneously when they can really only focus on one of them, or make a conscious effort to reduce usage time. Therefore, instead of simply providing occupants with a number of general energy-saving recommendations, it is more rational and efficient to help them modify their behavior in two steps. First, it is necessary to identify the behavior that needs to be modified. This can be achieved by analyzing measured building-related data. Second, feasible recommendations to mitigate the identified behavior can be presented with the goal of reducing energy consumption in the home.

With regard to the building simulation approach, current simulation tools can only imitate some typical activities in a rigid way, such as the control of sun-shading devices,

while realistic building occupant behavior patterns are more complicated.

In summary, a methodology is needed for evaluating occupant behavior in existing residential buildings and for helping occupants modify their activities efficiently through analyzing measured building-related data. In this research, a methodology is proposed based on the three data mining techniques: clustering analysis, data classification, and association rule mining.

3. DATA MINING PROCESS AND FRAMEWORK FOR KNOWLEDGE DISCOVERY

In this chapter, a data analysis process and a systematic data mining framework aimed at mining hidden and useful knowledge from measured building-related data are proposed. Three data mining techniques are introduced: data classification, cluster analysis, and association rule mining. Finally, measured building-related data collected for the case studies in this research is described.

3.1 Proposed Data Analysis Process

The adopted data analysis process is depicted in Figure 3-1.

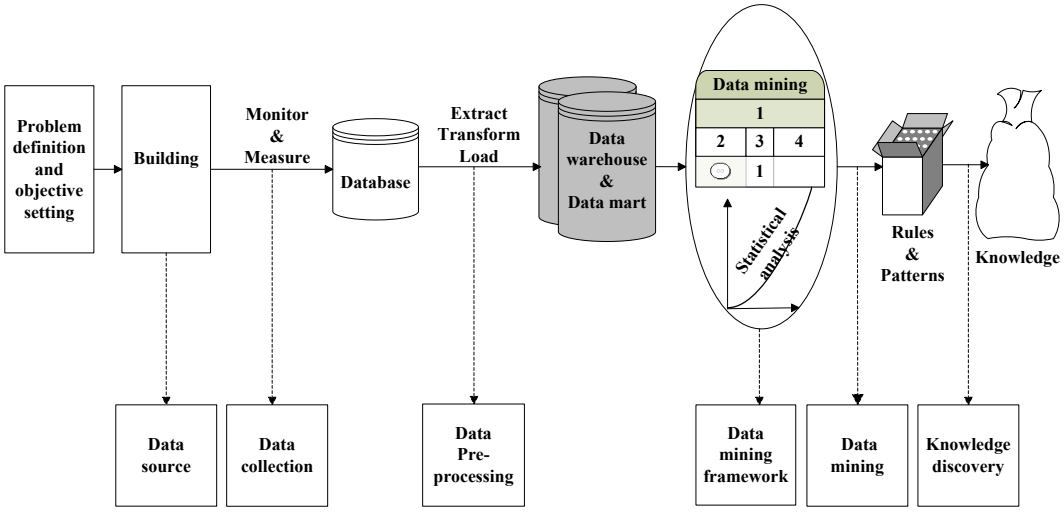

Figure 3-1 Process for data analysis within the building engineering domain

The process consists of the following steps:

(1) Problem definition and objective setting;

(2) Data source selection: select buildings available to collect measured building-related data;

(3) Data collection: collect building-related data through building automation systems, field survey, etc., and then construct a database;

(4) Data preprocessing/preparation: detect and remove outliers and noise, handle missing values, deal with inconsistencies and complexity through data transformation and integration, etc.;

(5) Data warehouses (DWs) or data marts construction: construct DWs or data marts so as to provide on-line analytical processing. The gray block in Figure 3-1 denotes that the step was unnecessary in this study. First, the measured building-related data were collected and processed while on-line analytical processing was not necessary. Second, the database is relatively small and there is no need to build a high-dimensional DW;

(6) Data mining and model construction: perform data mining based on the proposed data mining framework. In particular, three data mining techniques are utilized: data classification, cluster analysis, and association rule mining. Traditional statistical analysis is also employed as a supplementary tool, such as the verification of data mining results;

(7) Results analysis and evaluation: identify the most useful rules and patterns from the data mining results;

(8) Knowledge discovery and presentation: discover useful knowledge based on both

expertise and obtained rules/patterns.

3.2 Proposed Data Mining Framework

Figure 3-2 shows the data mining framework proposed in this study. The framework is composed of measured building-related data, selected data mining techniques and algorithms, and output of useful knowledge about building energy performance evaluation and improvement.

In this framework, three data mining techniques are employed as a primary tool. Note that each data mining technique can be achieved by different algorithms. For example, data classification can be conducted by using the decision/regression tree algorithm; cluster analysis can be conducted by using the *K*-Means/*K*-Modes algorithm; association rule mining can be conducted by using the Apriori/FP-growth algorithm. Furthermore, different data mining techniques can be combined to mine building-related data, such as the cluster analysis and data classification (e.g. clustering-then-classification, see Chapter 7), the cluster analysis and association rule mining (e.g. association rule clustering system and frequent pattern-based clustering analysis). For demonstration purposes, some algorithms (e.g. decision tree, *K*-means clustering, and FP-tree) were used in this study to address the four problems outlined in Chapter 1. An overview of these data mining algorithms is presented in the following sections. The reader can refer to the data mining textbooks (Cios, et al., 2007; Rokach and Maimon, 2008; Cao et al., 2009) for more detailed descriptions and mathematic formula of the algorithms.

Based on the proposed process and framework, architects and researchers could analyze measured building-related data efficiently and extract useful hidden knowledge which could help to account for interactions between building energy consumption and its influencing factors. Note that a clear and thorough understanding of such interactions could provide essential guidance in presenting energy-saving opportunities.

3.3 Data Mining Techniques

This section first present basic terms and concepts in relation to data mining, and then introduces the three data mining techniques, as well as the data mining algorithms employed in this study. Useful terminologies are:

- **Dataset, Attribute, and Instance**: a dataset is a set of data items. It is roughly equivalent to a two-dimensional (i.e. column and row) spreadsheet or database table, as shown in Figure 3-3. Each database table consists of a set of attributes (usually in different columns or fields) and stores a large set of instances (usually in rows or records). Consider an HVAC system with 100 monitored parameters. Each parameter can be considered an attribute, and a record of all these parameters in a specific time point can be considered an instance.

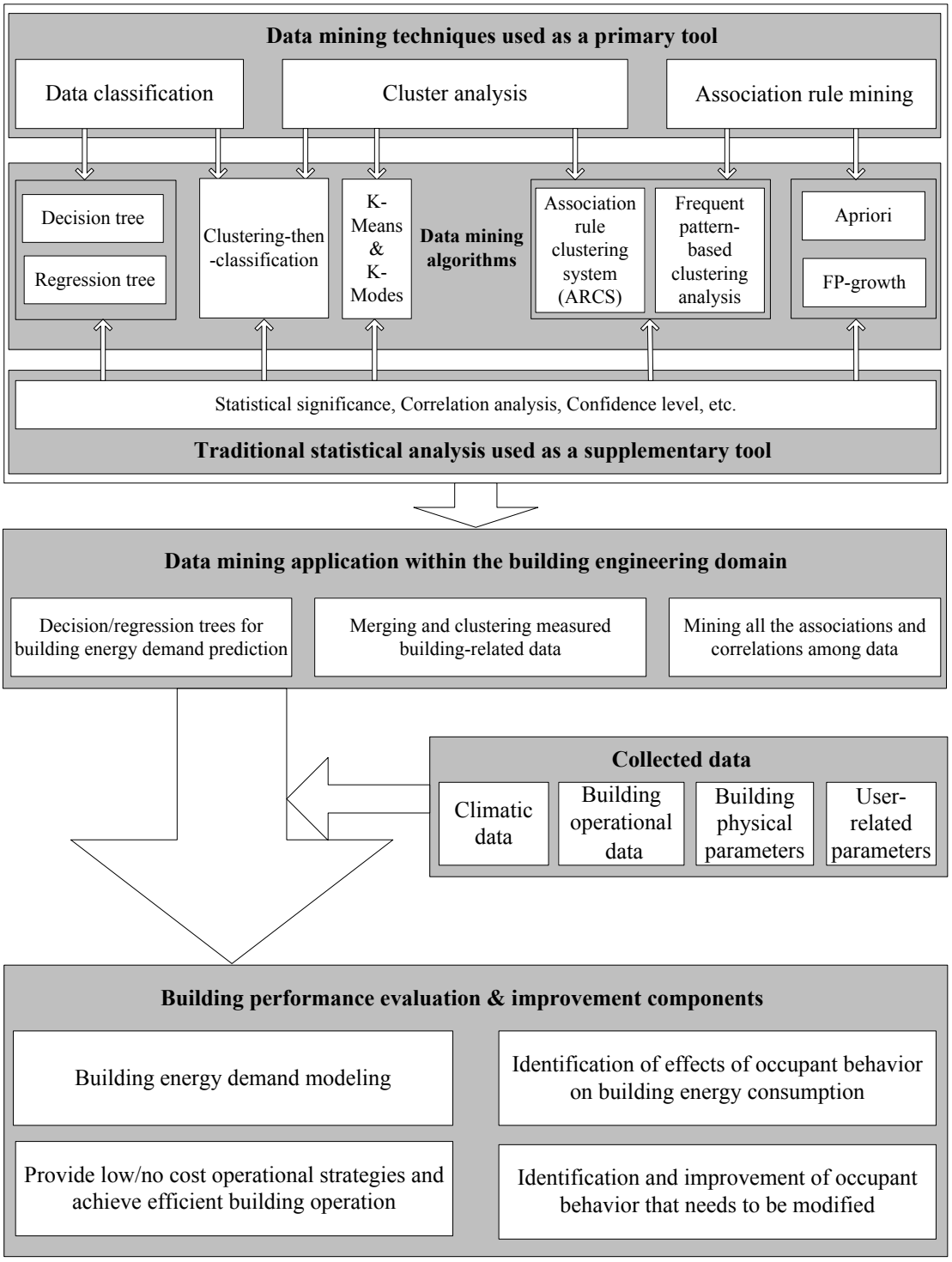

Figure 3-2 Overview of the proposed data mining framework

Attribute Instance	Attribute 1	...	Attribute m
Instance 1	x	...	x
…	...	...	...
Instance i	x	...	x
...	...	...	...
Instance j	x	...	x
...	...	...	...
Instance n	x	...	x

Figure 3-3 A schematic diagram of dataset, attribute and instance

- **Target attribute, Predictor attribute**: Target attribute is the attribute predicted as a function of other attributes (i.e. predictor attributes). For example, the building energy consumption is the target attribute, and could be predicted as a function of building-related parameters such as floor area and number of occupants (i.e. predictor attributes).

Based on the above explanation of the data mining terms, data classification, cluster analysis, and association rule mining, are described as follows.

3.3.1 Data Classification and Decision Tree

Overview of Decision Tree

The decision tree method is one of the most commonly used data mining methods (Quinlan, 1986; Han et al., 2006). It uses a flowchart-like tree structure to segregate a set of data into various predefined classes, thereby providing the description, categorization, and

generalization of given datasets. As a logical model, decision tree shows how the value of a *target variable* can be predicted by using the values of a set of *predictor variables*. Figure 3-4 presents a decision tree indicating whether residents turn room air conditioners (RAC) on or off in their rooms in the cooling season. For this example, assume 100 instances are used to build this decision tree, and that each instance has three attributes: outdoor air temperature, room occupancy, and the operating state of RAC.

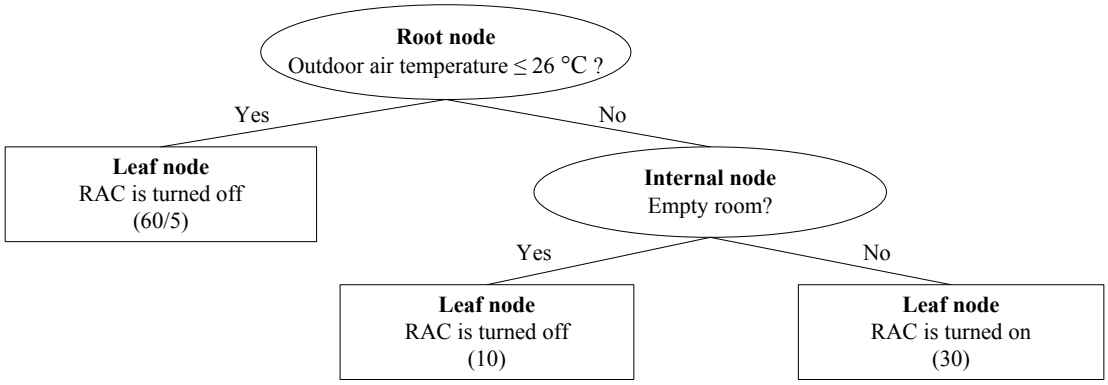

Figure 3-4 Schematic illustration of a simple hypothetical decision tree

The target variable for the above decision tree is RAC operating states, with potential states being classified as either turning on or off. The predictor variables are outdoor air temperature (≤ 26 °C or > 26°C) and room occupancy (empty or not). As shown in Figure 3-4, a decision tree consists of three kinds of nodes: root node, internal node, and leaf node. Root nodes and internal nodes denote a binary split test on an attribute while leaf nodes represent an outcome of the classification (i.e., a categorical target label). Moreover, the

numbers in the parentheses at the end of each leaf node depicts the number of instances in this leaf. If some leaf nodes are impure (i.e., some records are misclassified into this node), the number of misclassified instances will be given after a slash. For example, (60/5) in the left most leaf in Figure 3-4 means that, among the 60 instances having outdoor temperature is lower than or equal to 26 °C that have been classified to *turned off*, 5 of them actually have the value *turned on*. By using this decision tree, the RAC operating state classification (i.e. turn on or turn off) can be predicted. For example, if the outdoor air temperature is higher than 26 °C and the room is not empty, occupants will turn RAC on; otherwise, they will turn it off.

Decision tree generation is in general a two-step process, namely learning and classification, as shown in Figure 3-5. In the learning process, the collected data is split into two subsets, a training set and a testing set. Creation of the training and testing sets is an important part of evaluating data mining models. Usually, most of the instances in the database are arbitrarily selected for training and the remained instances are used for testing. Note that the training and testing sets should come from the same population but should be disjoint. Then, a decision tree generation algorithm takes the training data as an input, with the corresponding output being a decision tree. Commonly used decision tree generation algorithms include ID3 (Quinlan, 1986), classification and regression trees (CART) (Breiman et al., 1984), and C4.5 (Quinlan, 1993). In the classification process, the accuracy of the obtained decision tree is first evaluated by making predictions against test

data. The accuracy of a decision tree is measured by comparing the predicted target values with the true target values of the test data. If the accuracy is considered acceptable, the decision tree can be applied to new dataset for classification and prediction; otherwise, the reason for any inaccuracies should be identified and corresponding solutions should be adopted to address these problems.

Decision Tree Generation

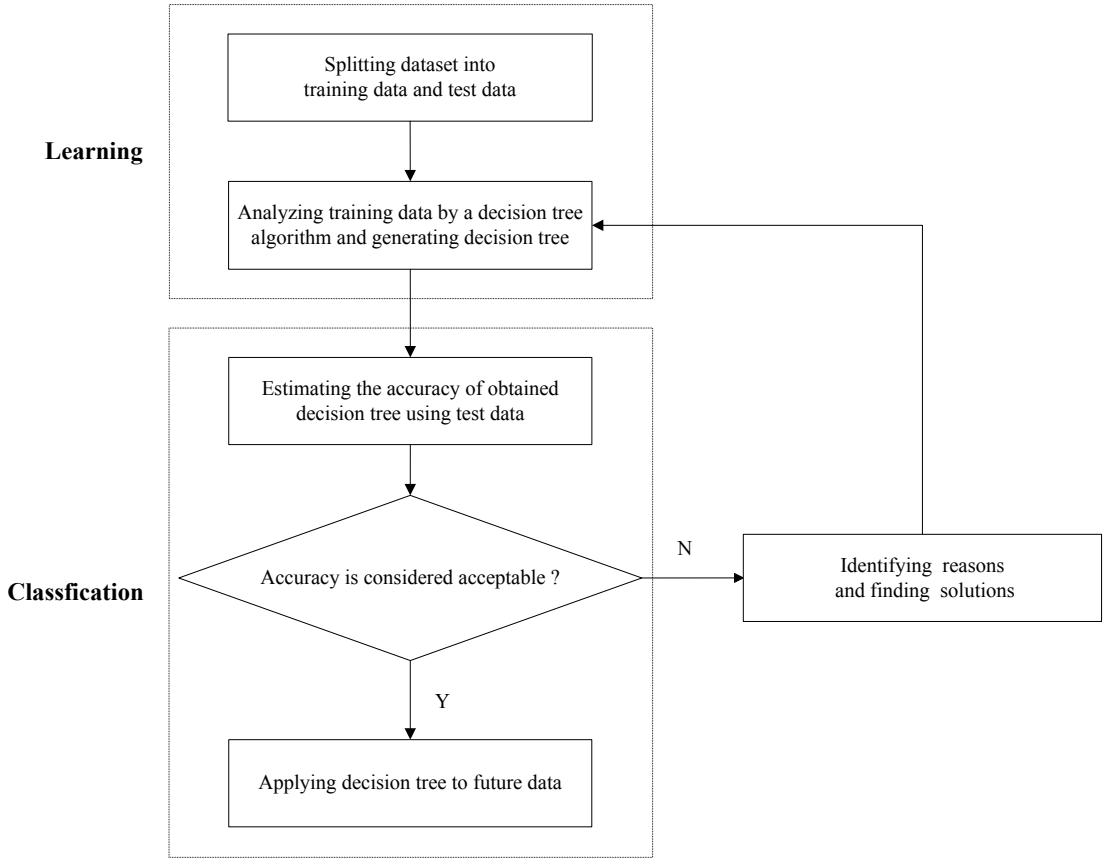

Figure 3-5 Procedure of decision tree generation

The procedure for generating a decision tree from the training data is explained as follows. Initially, all instances in the training data are grouped together into a single partition. At each iteration the algorithm chooses a predictor attribute that can "best" separate the target class values in the partition. The ability of a predictor attribute to separate the target class values is measured based on an attribute selection criterion, which will be discussed in the following section. After a predictor attribute is chosen, the algorithm splits the partition into child partitions such that each child partition contains the same value of the chosen selected attribute. The decision tree algorithm iteratively splits a partition and stops when any one of the following terminating conditions is met:

(1) All instances in a partition share the same target class value. Thus, the class label of the leaf node is the target class value.

(2) There are no remaining predictor attributes that can be used to further split a partition. In this case, the majority target class values becomes the label of the leaf node.

(3) There are no more instances for a particular value of a predictor variable. In this case, a lead node is created with the majority class value in the parent partition.

Attribute Selection Criterion

The decision tree generation algorithm is a greedy algorithm. It iteratively splits a partition by choosing a split attribute that can best separate the target class values. The choice of split attribute determines the quality of the decision tree model and, therefore, the classification accuracy on the future data. The concept of *entropy* (Han et al., 2006) in

information theory is a widely used criterion measure for decision tree to characterize the purity of a partition in decision tree nodes. Assume a decision tree containing only binary target variables (e.g. HIGH and LOW), the entropy of the data subset, D_i, of the ith tree node is defined as

$$Entropy(D_i) = -(\frac{n_{HIGH}}{T\_N} log_2 \frac{n_{HIGH}}{T\_N} + \frac{n_{LOW}}{T\_N} log_2 \frac{n_{LOW}}{T\_N}) \quad [3.1]$$

where

n_{HIGH}: the number of HIGH EUI records in D_i

n_{LOW}: the number of LOW EUI records in D_i

T_N: the total number of records in D_i and $T\_N = n_{\text{HIGH}} + n_{\text{LOW}}$

The entropy varies between 0 and 1. Notice that the entropy equals to 0 if D_i is pure and it is 1 when n_{HIGH} equals to n_{LOW}. At each node of a decision tree, candidate splitting test is used to evaluate all available attributes to select the most suitable attribute to split data. Suppose the jth attribute has been selected as node attribute. A candidate split test, ST, at the ith tree node is defined as

$$ST: Val_j(r) \leq T\_h \quad \text{(if the } j\text{th attribute is numerical)} \quad [3.2]$$

$$ST: Val_j(r) \in \{v_1, v_2\} \quad \text{(if the } j\text{th attribute is categorical and has two values)} \quad [3.3]$$

where

$Val_j(r)$: the value of the jth attribute of record r

T_h: threshold value

v_1, v_2: two values of the jth attribute

Next, the algorithm applies *ST* to D_i and partitions it into two subsets, DS_1 and DS_2. Let *r* be a record in D_i. If the *j*th attribute is a numerical attribute, then

$$DS_1 = \{r \in D_i | Val_j(r) \leq T\_h\} \text{ and } DS_2 = \{r \in D_i | Val_j(r) > T\_h\} \quad [3.4]$$

If the *j*th attribute is a categorical attribute, then

$$DS_1 = \{r \in D_i | Val_j(r) = v_1\} \text{ and } DS_2 = \{r \in D_i | Val_j(r) = v_2\} \quad [3.5]$$

Let *m* and *n* be the numbers of instances in DS_1 and DS_2, respectively. The entropy after the split test can then be calculated as the weighted sum of the entropies for the individual subsets

$$Entropy(DS_1 \& DS_2) = \frac{m}{m+n} Entropy(DS_1) + \frac{n}{m+n} Entropy(DS_2) \quad [3.6]$$

The selection of node attributes used to split data is important and a rational selection can improve the purity of tree nodes. A widely used attribute selection measure is *information gain* (Shannon, 1948), which is defined as the entropy reduction before and after a candidate splitting test. Therefore, information gain can be calculated as

$$InfoGain = Entropy(D_i) - Entropy(DS_1 \& DS_2) \quad [3.7]$$

For each tree node, the attribute with the maximum information gain will be chosen as the splitting attribute at this node. The information gain measure, however, has a bias to attributes with larger number of domain values. One way to avoid such a bias is to normalize the information gain by a split information value defined analogously with information gain. C4.5 algorithm employs this improved measure, *gain ratio* (Han et al., 2006):

$$GainRatio = \frac{InfoGain}{SplitInfo} \quad [3.8]$$

where

$$SplitInfo = -(\frac{m}{m+n}\log_2\frac{m}{m+n} + \frac{n}{m+n}\log_2\frac{n}{m+n}) \quad [3.9]$$

The attribute with the highest gain ratio is selected as the splitting attribute.

Additionally, in order to detect leaf nodes, a minimum threshold value of entropy (EN_{min})is predefined and compared with node classification entropy ($Entropy(D_i)$), if $Entropy(D_i)$ is lower than EN_{min}, then this node is a leaf and will be labeled LEAF. Otherwise a further splitting test should be performed. However, if no significant effects are observed on information gain or gain ratio in further candidate splitting tests, the test will be also stopped and the node will be labeled STOP.

3.3.2 Cluster Analysis and the *K*-means Algorithm

Cluster analysis is the process of grouping the observations into classes or clusters so that objects in the same cluster have a high similarity, while objects in different clusters have a low similarity. Figure 3-6 shows a clustering schema based on a hypothetical building data table. It contains various energy-related variables such as outdoor air temperature (T) and building heat loss coefficient (HLC).

The data table consists of *m* attributes and *n* instances. Each attribute represents a variable and each instance denotes a building. All the instances are grouped into *w* clusters. These *w* clusters are homogeneous internally and heterogeneous between different clusters

(Han et al., 2006). Such internal cohesion and external separation are based upon the *m* attributes; it implies that these attributes have the most similar holistic effects on the building energy performance of the same cluster buildings, while the effects are significantly distinct for the buildings in different clusters.

	Instance	Attribute 1 (T)	...	Attribute m (HLC)
Cluster 1	Instance 1	x	x	x
	…	x	x	x
	Instance i	x	x	x
⋮	...	x	x	x
	Instance j	x	x	x
Cluster w	...	x	x	x
	Instance n	x	x	x

Figure 3-6 Clustering schema

The dissimilarity between observations in the database is calculated using the distance between them in the cluster analysis. In this study, the most commonly used distance measure, Euclidean distance, was used (Han et al., 2006):

$$d(k,l) = \sqrt{(x_{k1} - x_{l1})^2 + (x_{k2} - x_{l2})^2 + \cdots + (x_{kn} - x_{ln})^2} \quad [3.10]$$

where $k = (x_{k1}, x_{k2}, \ldots, x_{kn})$ and $l = (x_{l1}, x_{l2}, \ldots, x_{ln})$ are buildings. $x_{k1}, \ldots, x_{kn}$ are n parameters of k and $x_{l1}, \ldots, x_{ln}$ are n parameters of l.

Commonly used clustering algorithms include *K*-means, *K*-medoids, and CLARANS (Han et al., 2006). In this study, we employed the *K*-means, along with the open-source

data mining program RapidMiner (Rapid-I 2001), to perform cluster analysis due to its high efficiency and wide applicability.

The *K*-means algorithm is one of the simplest partition methods to solve clustering problem. Given a dataset (*D*) containing *w* objects, the *K*-means algorithm aims to partition these *w* objects into *k* clusters with two restraints: 1) the center of each cluster is the mean position of all objects in that cluster, 2) each object has been assigned to the cluster with the closest center. This algorithm consists of given steps: 1) Randomly select *k* observations from *D* as the initial cluster centers, 2) Calculate the distance between each remaining observation and each initially chosen center, 3) Assign each remaining observation to the cluster with the closest center, 4) Recalculate the mean values, i.e., the cluster centers, of the new clusters, and 5) Repeat Steps 2 to 4 until the algorithm converges, meaning that the cluster centers do not change.

In RapidMiner, the performance of a clustering algorithm is evaluated by the Davies Bouldin index (DBI) (Davies and Bouldin, 1979). This index is defined as the ratio of the sum of average distance inside clusters to distance between clusters.

$$DBI = \frac{1}{n}\sum_{i=1}^{n} max_{i \neq j}\left[\frac{R_i + R_j}{M_{i,j}}\right] \qquad [3.11]$$

where

n: number of clusters

R_i, R_j: average distance inside cluster i and cluster j by averaging the distance between each cluster object and the cluster center

$M_{i,j}$: distance between cluster centers

The DBI is small if each cluster is comparatively dense; while different clusters are far from each other. Consequently, a smaller DBI indicates better performance.

It should be mentioned that the *K*-means is sensitive to initial cluster centers. Therefore, different values should be tried so as to obtain the minimum sum of the distances within a cluster. At the same time, the number of clusters should be specified in advance.

3.3.3 Association Rule Mining

In data mining, association rules are often used to represent patterns of parameters that are frequently associated together. An example is given to illustrate the concept of association rules. Assume that 100 occupants live in 100 different rooms in the same building and each room has both a window and a door. Moreover, 40 occupants open the windows and 20 occupants open the doors. If 10 occupants open both the windows and doors simultaneously, it can be calculated that these 10 occupants account for 10% of all the building occupants (10/100 = 10%), and 25% of the occupants who open windows (10/40 = 25%). Then, the information that occupants who open windows also tend to open doors at the same time can be represented in the following association rule:

$$\text{open\_windows} \rightarrow \text{open\_doors}\ [support = 10\%, confidence = 25\%]$$

In this statement, *support* and *confidence* are employed to indicate the validity and certainty of this association rule. Different users or domain experts can set different thresholds for *support* and *confidence* according to their own requirements, in order to discover useful knowledge eventually. Accordingly, the association rule mining (ARM) can be defined as finding out association rules that satisfy the predefined minimum *support* and *confidence* from a given database.

Mathematically, *support* and *confidence* can be calculated by probability, $P(X \cup Y)$, and conditional probability, $P(Y|X)$, respectively (X denotes the premise and Y denotes the consequence in the sequence). That is,

$$support(X \rightarrow Y) = P(X \cup Y) \quad [3.12]$$

$$confident(X \rightarrow Y) = P(Y|X) \quad [3.13]$$

Another concept, *lift,* which is similar to *confidence,* is commonly used to demonstrate the correlation between the occurrence of X and Y when conducting the ARM. Mathematically,

$$lift(X \rightarrow Y) = \frac{P(X \cup Y)}{P(X)P(Y)} = \frac{P(Y|X)}{P(Y)} \quad [3.14]$$

Particularly, a *lift* value greater than 1 represents a positive correlation (the higher this value is, the more likely that X coexists with Y, and there is a certain relationship between X and Y (Han et al., 2006) while a *lift* value less than 1 represents a negative correlation. If the value is equal to 1, i.e.$P(X \cup Y) = P(X)P(Y)$, the occurrence of X is independent of the occurrence of Y, and there is no correlation between X and Y.

Commonly used ARM algorithms include the Apriori algorithm and the frequent-pattern growth (FP-growth) algorithm (Han et al., 2006). In this study, the FP-growth algorithm was utilized due to its high efficiency and wide applicability. The specific algorithm of the FP-growth is presented in (Han et al., 2006).

3.4 Data Collection

To demonstrate the applicability of the proposed process and framework, the measured data from a set of Japanese residential buildings and from the EV building located in Montreal was collected and analyzed.

3.4.1 Measured Data from Residential Buildings

To evaluate and improve the energy performance of residential buildings, a project entitled "Investigation on Energy Consumption of Residents All over Japan" was carried out by the Architecture Institute of Japan from December 2002 to November 2004 (Murakami et al., 2006). For this project, field surveys on energy-related data and other relevant information were carried out in 80 residential buildings located in six different districts in Japan: Hokkaido, Tohoku, Hokuriku, Kanto, Kansai, and Kyushu. Table 3-1 shows the survey items and corresponding investigation methods. Figure 3-7 shows the measuring instruments which were used to monitor temperature and consumptions of electricity, gas, and/or kerosene.

Table 3-1 Investigation items and methods

Method	Survey items		Measuring time
Field measurement	Different end-use loads of all kinds of fuel	Electricity	Measured every minute
		Gas	Measured every 5 minutes
		Kerosene	Measured every 5 minutes
	Indoor air temperature (1.1m above floor)		Measured every 15 minutes
Questionnaire survey	Lifestyle, Utilization of equipment, Annual income, etc.		Once only
Inquiring survey	Other issues, such as basic building information		Once only

Figure 3-7 Measuring instruments (from left to right: electricity, gas, kerosene and air temperature)

The building energy consumption was broken down into eight major end-use loads: 1) HVAC, 2) supply hot water (SHW), 3) kitchen (KITC, including cooking and other kitchen equipment such as dishwasher and range hood), 4) lighting (LIGHT), 5) refrigerator (REF), 6) amusement and information (A&I, such as television, telephone, and computer, etc.), 7) housework and sanitary (H&S, such as washing machine, vacuum, and electrical shaver, etc.), and 8) others (OTHER, unidentified usage such as electrical shutter and all the unclear items).

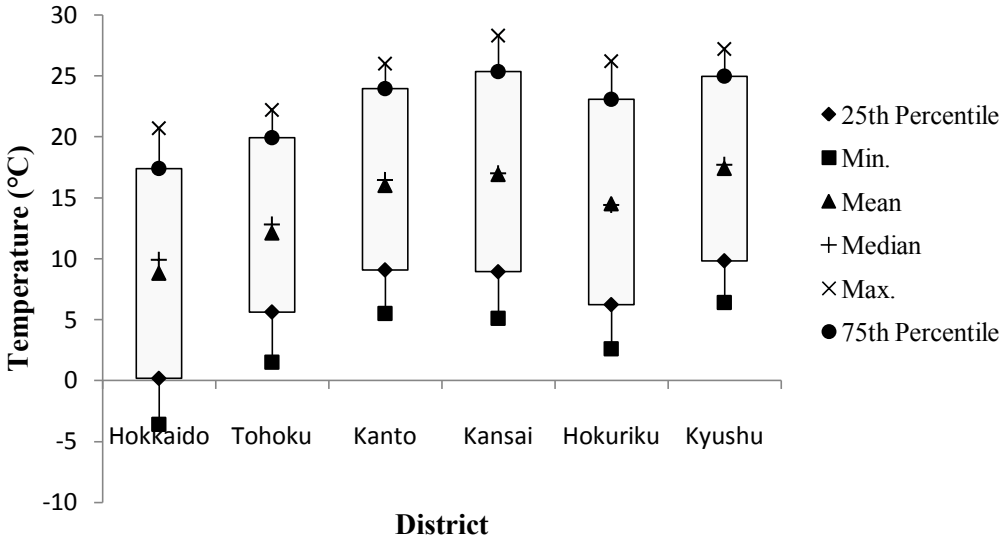

Figure 3-8 Boxplot for monthly average outdoor temperature in 2003

Figure 3-8 shows the boxplot for monthly average outdoor air temperature in each district in 2003 using Japanese meteorological data. The mean value of monthly average temperature, i.e., annual average temperature, is also given. Clearly the monthly average temperature has a more or less symmetric distribution. The annual average temperature is higher than 8 °C in all the six districts and the temperature in Hokkaido and Tohoku is comparatively lower than other districts.

Scrutinizing the data from the 80 buildings it was found that only 67 sets were complete while the other 13 had missing values of energy consumption data. Figure 3-9 shows the percentage breakdown of available residential buildings in each district. It can be seen that

the distribution is roughly uniform.

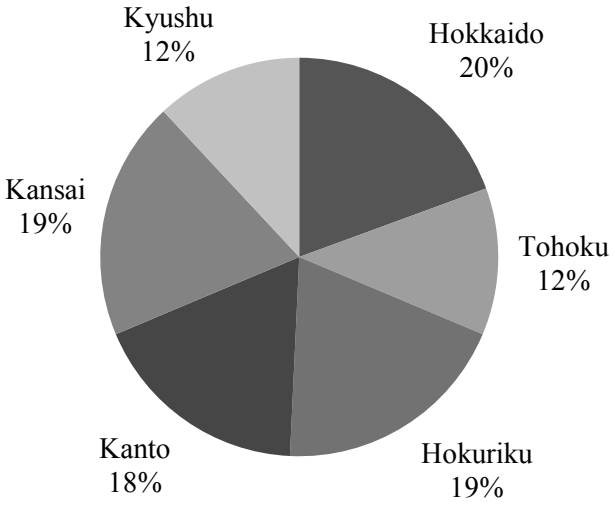

Figure 3-9 Percentage breakdown of buildings in each district

Table 3-2 Conversion coefficients of different fuels

Fuel	Conversion coefficient	Unit
Electricity	3.6	MJ/kWh
City gas (4A-7C)	20.4	MJ/Nm^3
City gas (12A-13C)	45.9	MJ/Nm^3
Liquefied petroleum gas (LPG)	50.2	MJ/Nm^3
Kerosene	36.7	MJ/L

Data reduction and aggregation was then performed to obtain a smaller representation of the original data. For example, diverse energy unit of different kinds of primary energy sources used by various buildings, including electricity, natural gas, and kerosene, was converted to MJ based on conversion coefficients given in Table 3-2. Then, readings of each end-use load at different intervals (e.g. 1 or 5 minutes) were averaged over each month. The resulting data was stored in a database.

3.4.2 Measured Data from the EV Building

The EV pavilion located in Montreal, a complex building that mainly includes offices and chemical labs, was selected as data source for this study. This building consists of two parts: the ENCS part (17 floors) and the VA part (12 floors), as shown in Figure 3-10.

Figure 3-10 EV Pavilion

Each pavilion has its own VAV air-conditioning system. In the ENCS part, the air handling units (AHUs) are installed in the local mechanic rooms on each floor except for the 17th floor (the mechanical floor), where various equipment, such as the chillers and fresh air handling units (FHUs), are installed. On the 17th floor, two identical FHUs (i.e., the FHU 1 and FHU 2) are employed to process fresh air and each has two variable speed fans in parallel, as shown in Figure 3-11. Due to the existence of chemical labs in the ENCS part, the fresh air is separated into two streams: stream 1 is sent to the local mechanical rooms in each floor and mixed with the return air from the same floor's

rooms other than chemical labs. Then the mixed air is conditioned by the AHUs in that floor's mechanical room and supplied to those rooms again. Meanwhile, stream 2 is mixed with the return air from the atriums in the ENCS part. Then the mixed air is conditioned by the FHU 3, which also has two variable speed fans in parallel, and sent to the chemical labs. The exhaust air from both the chemical labs and other rooms is discharged to outside directly by the EHU 1, which contains two variable speed fans, as shown in the dash line square. Moreover, the dash dot line in Figure 3-11 indicates a recuperation loop installed between the fresh air and the exhaust air to exchange heat in both cooling and heating seasons.

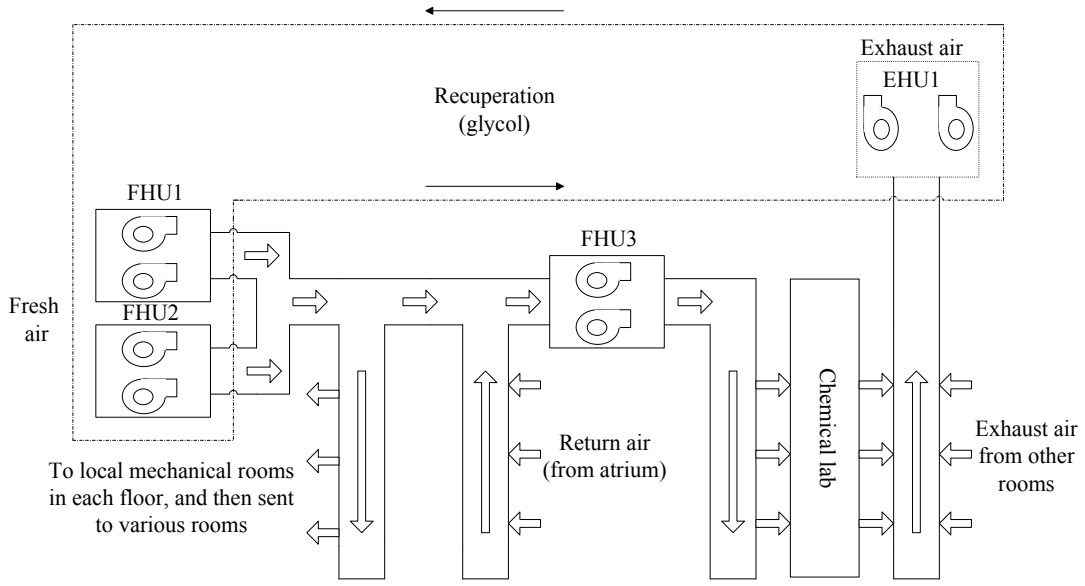

Figure 3-11 Flow chart of air-conditioning system in the ENCS pavilion

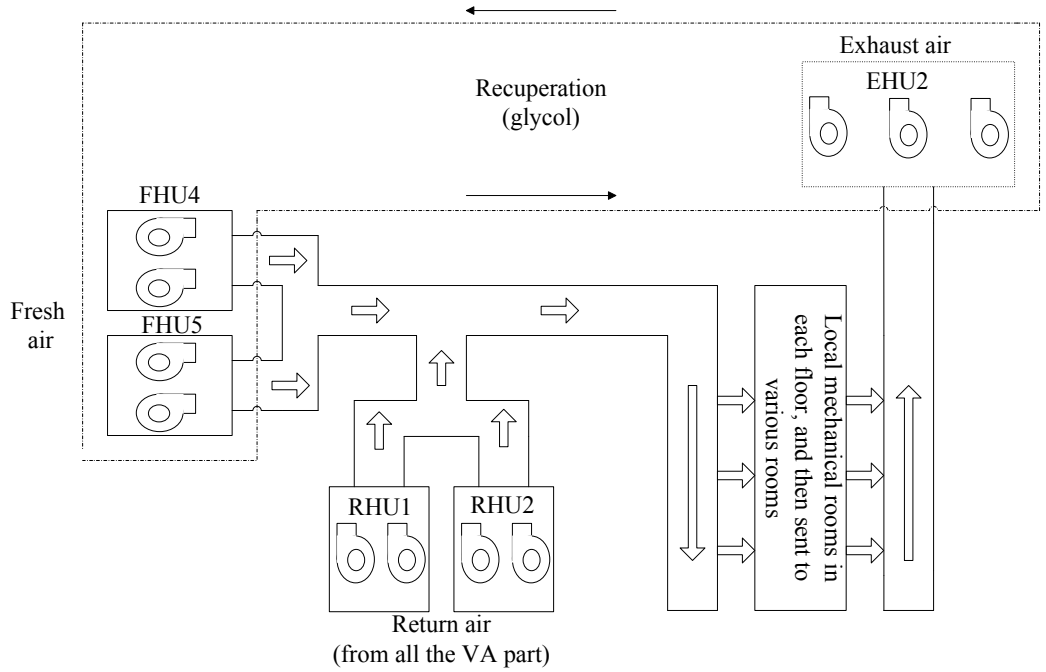

Figure 3-12 Flow chart of air-conditioning system in the VA pavilion

The flow chart of air-conditioning system in the VA pavilion is shown in Figure 3-12. Similarly, the air handling units (AHUs) are installed in the local mechanical rooms on each floor except for the 12th floor (the mechanical floor), where various equipment, such as the chillers and fresh air handling units, are installed. On the 12th floor, two identical FHUs (i.e., the FHU 4 and FHU 5) are employed to process fresh air and each of them has two variable speed fans in parallel. Given that there is no chemical lab in the VA pavilion, the fresh air is mixed with the return air from all the VA part directly. The mixed air is sent to the local mechanical rooms in each floor to be conditioned by the AHUs, and then sent to various rooms in the same floor. Two RHUs (i.e., the RHU 1 and RHU 2) are

employed to return air, and each of them has two variable speed fans in parallel. The exhaust air in the VA pavilion is discharged to outside by the EHU 2, which contains three variable speed fans, as shown by dash line square. Also, the dash dot line in Figure 3-12 indicates a recuperation loop installed between the fresh air and exhaust air to exchange heat in both cooling and heating seasons.

Table 3-3 The monitored parameters of the air-conditioning systems

No.	Parameter	No.	Parameter	No.	Parameter	No.	Parameter
1	Q_{I1}	21	TA_{IVac}	41	TA_{IXari}	61	F_{IX3}
2	Q_{I2}	22	TA_{Vac}	42	TA_{IXarii}		
3	Q_{II1}	23	TA_{Iah}	43	TA_{IXariii}		
4	Q_{II2}	24	TA_{IIah}	44	TA_{oENCS}		
5	Q_{III1}	25	TA_{IVah}	45	H_{oENCS}		
6	Q_{III2}	26	TA_{Vah}	46	TA_{oVA}		
7	Q_{IV1}	27	TA_{IIbr}	47	H_{oVA}		
8	Q_{IV2}	28	TA_{IVbr}	48	TG_{ENCSar}		
9	Q_{V1}	29	TA_{Vbr}	49	TG_{VAar}		
10	Q_{V2}	30	TA_{Iar}	50	F_{I}		
11	Q_{III}	31	TA_{IIar}	51	F_{II}		
12	Q_{VI}	32	TA_{IVar}	52	F_{III}		
13	Q_{VII}	33	TA_{Var}	53	F_{IV}		
14	Q_{VIII1}	34	TA_{VIIIbri}	54	F_{V}		
15	Q_{VIII2}	35	TA_{VIIIbrii}	55	F_{VI}		
16	Q_{VIII3}	36	TA_{IXbri}	56	F_{VII}		
17	Q_{IX1}	37	TA_{IXbrii}	57	F_{VIII1}		
18	Q_{IX2}	38	TA_{IXbriii}	58	F_{VIII2}		
19	TA_{Iac}	39	TA_{VIIIari}	59	F_{IX1}		
20	TA_{IIac}	40	TA_{VIIIarii}	60	F_{IX2}		

In order to conduct the case study, the historical data of the air-conditioning systems

in both parts were collected from December 2006 to May 2009. However, since the online monitoring program was updated from November 2007 to January 2008, data reports were not generated during this period. In total, 61 parameters were monitored in the two air-conditioning systems and data of each parameter was trended at a 15-minute interval. The monitored parameters are given in Table 3-3.

4. A DECISION TREE METHOD FOR BUILDING ENERGY DEMAND MODELING

4.1 Introduction

In the design of an energy efficient building, architects and building designers often need to identify which parameters influence future building energy demand significantly. Furthermore, based on different combinations of these parameters as well as their values, architects and building designers usually expect to find a simple and reliable method to estimate building energy performance rapidly so that they can optimize their building design plans. In recent years, there have been many studies on building energy demand modeling, and several methods were employed, mainly including traditional regression methods and artificial neural networks (ANN) methods. Both of the methods have been reviewed in Chapter 2; and in this study the decision tree method is proposed to remove their limitations.

In the past two decades, the decision tree method, a novel computational modeling technique that uses flowchart-like tree structure, was widely used for classification and prediction in many scientific and medical fields. The popularity of the decision tree method can be attributed to its ease of use, and abilities to generate accurate predictive models with understandable and interpretable structures, which, accordingly, provide clear and useful knowledge of corresponding domains. Moreover, the decision tree method is able to process both numerical and categorical variables, and perform classification and prediction

tasks rapidly without requiring much computation efforts. However, it should be mentioned that the decision tree method, as a classification analysis method, is more appropriate for predicting categorical variables than for predicting numerical variables[1]. The application of the decision tree method in building-related studies is still very sparse.

4.2 Methodology, Model target/input variables

4.2.1 Methodology

Figure 4-1 shows the methodology proposed for building energy demand modeling based on the decision tree method. Data collection and data pre-processing are first conducted. Then the learning process and the classification process are performed in turn. These two processes were described in detail in Chapter 3. If the accuracy of generated decision trees is considered acceptable, the decision trees can be applied to energy demand modeling. Commonly used decision tree generation algorithms include ID3, classification and regression trees (CART), and C4.5. In this study, we employ C4.5 algorithm, along with the open-source data mining software WEKA (Bouckaert et al., 2009). This software is selected due to its flexibility and wide applicability to different types of data.

[1] In the data mining filed, data classification differs from data prediction. In classification analysis, the target attribute is a categorical attribute. In prediction, the target attribute is a numerical attribute.

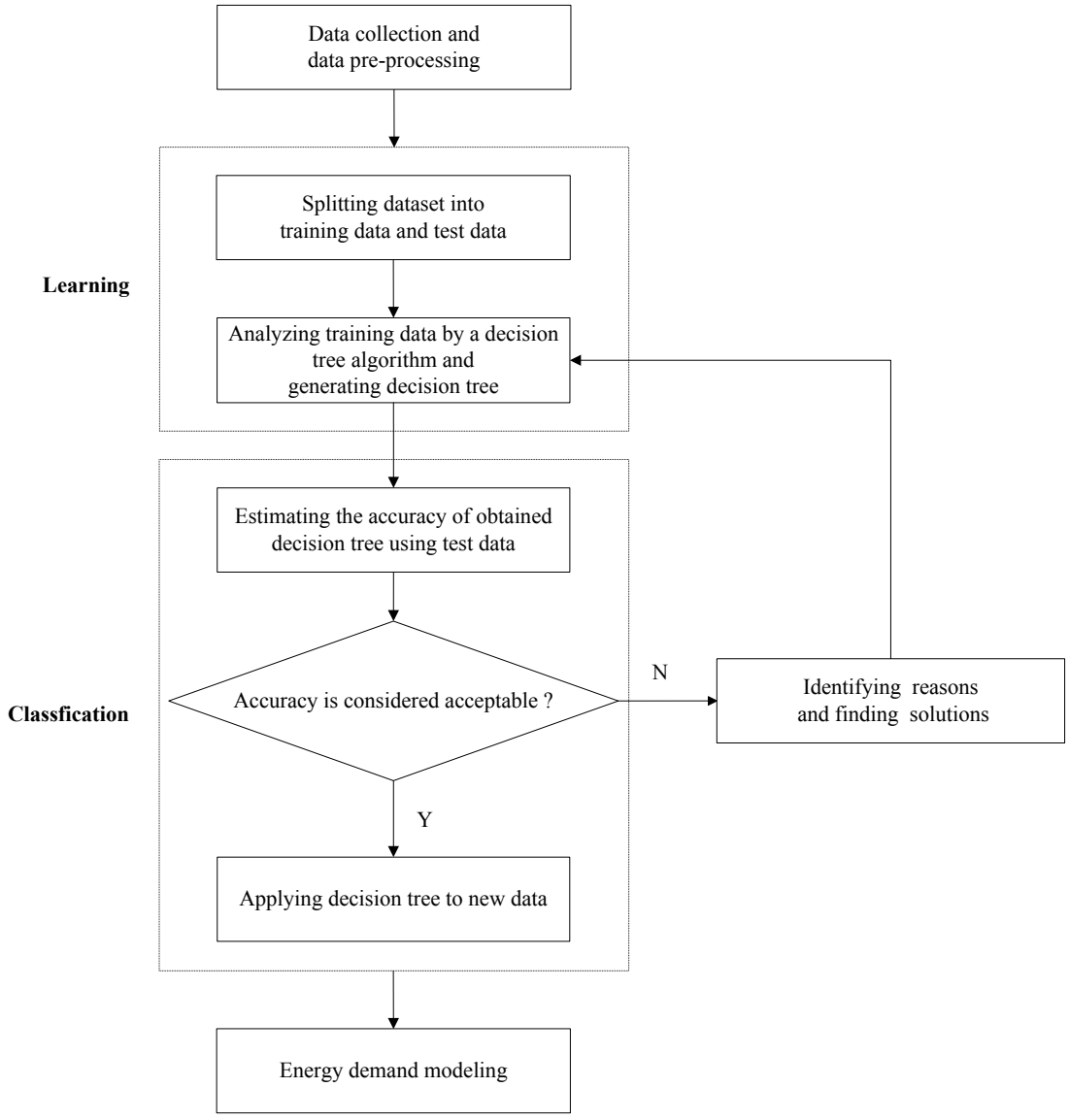

Figure 4-1 Proposed methodology for building energy demand modeling

The applicability of the methodology is demonstrated for the Japanese residential buildings.

4.2.2 Model target variable

In order to demonstrate building energy performance, model target variable is expressed in energy use intensity (EUI), defined as the ratio of annual total energy use to the total floor area (the annual total energy use is calculated as the sum of the energy content of all fuel used). As mentioned previously, the decision tree method is more appropriate for predicting categorical variables. Therefore, a concept hierarchy for building EUI is formed prior to the classification and prediction. Due to the small database size, a two grade descending scale, i.e., high level and low level, corresponding to low energy performance and high energy performance, are considered applicable and understandable. Building EUI ranges from 176 MJ/m^2 to 707 MJ/m^2 in the database and thus data ranged from the average of the maximum and minimum to the maximum value, [441.5, 707], is considered 'HIGH'. And data from the minimum value to the average of the maximum and minimum, [176, 441.5), is considered 'LOW'.

It should be mentioned that decision tree can also be used to classify and predict multiple EUI levels rather than just two. For example, instead of 'HIGH' and 'LOW', a concept hierarchy of EUI may map real EUI values into four conceptual levels such as *EXCELLENT, GOOD, FAIR*, and *COMMON,* thereby resulting in a smaller data range of each level and providing a more detailed description. However, more conceptual levels require a larger database and may be prone to higher misclassification rate of data records and thus reduce the accuracy of decision tree models.

4.2.3 Model input variables

Ten parameters (or *attributes*) are selected from the database to be model input parameters and are summarized in Table 4-1.

Table 4-1 Summary of model input parameters

Number	Variable	Type	Value	Variable label (unit)
1	T	Categorical	High/Low	Annual average air temperature
2	HT	Categorical	Detached/Apartment	House type
3	CO	Categorical	Wood/Non-wood	Construction type
4	FA	Numerical	[70, 240]	Floor area (m^2)
5	HLC[a*]	Numerical	[1.01, 4.35]	Heat loss coefficient (W/m^2K)
6	ELA[b*]	Numerical	[0.35, 13.30]	Equivalent leakage area (cm^2/m^2)
7	NO	Numerical	[2, 6]	Number of occupants
8	HEAT	Categorical	Electric/Non-electric	Space heating
9	HWS	Categorical	Electric/Non-electric	Hot water supply
10	KITC	Categorical	Electric/Gas	Kitchen

a* Calculated based on building design plans.

b* Measured by the fan pressurization method.

These ten parameters are grouped into four categories that are important determinants of household energy demand.

(1) Climatic conditions (T). The range of annual average outdoor air temperature in the six districts is discretized into two intervals based on the same concept hierarchy as the EUI mentioned earlier: the low interval [8.8 °C, 14.3 °C], and the high interval (14.3 °C, 17.4 °C]. According to this discretization criterion, the low temperature districts include Hokkaido and Tohoku while the other four districts belong to high temperature districts,

(2) Building characteristics (HT, CO, FA, HLC, ELA). For building construction type, the non-wood type includes steel reinforced concrete (SRC), reinforced concrete (RC), and steel structure (S),

(3) Household characteristics (NO), and

(4) Household appliance energy sources (HEAT, HWS, KITC). Energy sources are divided into energy generated from electricity consumption and energy generated from other fuels such as kerosene and natural gas.

Figure 4-2 shows the distribution of all the categorical parameters. It can be observed that all the percentages range from 30% to 70%, indicating a fairly uniform distribution.

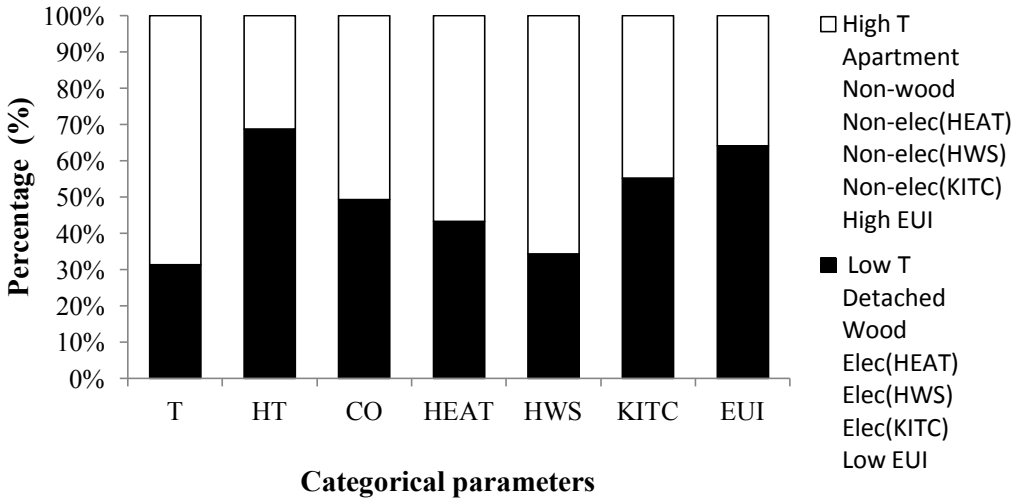

Figure 4-2 Categorical distribution of the six categorical parameters

4.3 Results and discussion

C4.5 algorithm was used for training data set (55 records were arbitrarily selected from the database) and test data set (the remaining 12 records) by using WEKA to build a decision tree for predicting the EUI of residential buildings to either 'HIGH' or 'LOW'.

4.3.1 Generation of decision tree

Figure 4-3 shows the decision tree for the classification of building EUI levels. This decision tree is built on the basis of the training data set of 55 data records with the ten attributes listed in Table 4-1. This tree includes a total of 21 nodes among which 11 are leaf nodes, including 8 LEAFs and 3 STOPs: this represents 11 classes (either EUI = HIGH or EUI = LOW). The explanatory note of three kinds of nodes, namely root node, internal node, and leaf node in this decision tree is shown in Figure 4-4. Note that entropy is also calculated and given in each node to characterize the purity of the sub dataset in that node. Moreover, the average EUI value of data records in each class is given and used for reference when performing prediction. Specifically, this reference value can be viewed as predictive numerical EUI value of the new data records that fall into that class.

The WEKA analysis report also provides the information on the classification accuracy of the decision tree. The report indicates that 51 records which accounts for 93% of all the training records are correctly classified: this indicates a good accuracy. Also, confusion matrix reports how many data records are correctly classified and misclassified in the class of HIGH EUI and LOW EUI separately, as below:

a b <-- classified as

35 1 | a = 'LOW EUI'

3 16 | b = 'HIGH EUI'

In this matrix, the number of correctly classified records is given in the main diagonal, i.e., upper-left to lower-right diagonal; the others are incorrectly classified. Only one instance of class "LOW EUI" was misclassified as "HIGH EUI" and three instances of class "HIGH EUI" was misclassified as "LOW EUI". Such information indicates that high EUI is more prone to be misclassified than low EUI. This may have occurred due to the fact that most of the data records are in LOW EUI so the tree became more sensitive to this class. An even distribution between the HIGH EUI and LOW EUI classes in database would possibly help to obtain sufficient accuracy and sensitivity in the desired classes.

The major strength of a decision tree lies in its interpretability and ease of use, particularly when decision rules are created. Based on a decision tree, decision rules can be easily generated by traversing a path from the root node to a leaf node. For example, a decision rule can be generated from Node 1 to Node 5 in above decision tree as follows: If T is high and HLC $\leq$ 3.89 and ELA $\leq$ 4.41 and HWS is electric then EUI is LOW. Since each leaf node produces a decision rule, the complete set of decision rules, which is equivalent to the decision tree, can be derived after all the leaf nodes have been included. Accordingly, above decision tree is converted to a set of decision rules, as show in Table 4-2.

4.3.2 Evaluation of the decision tree

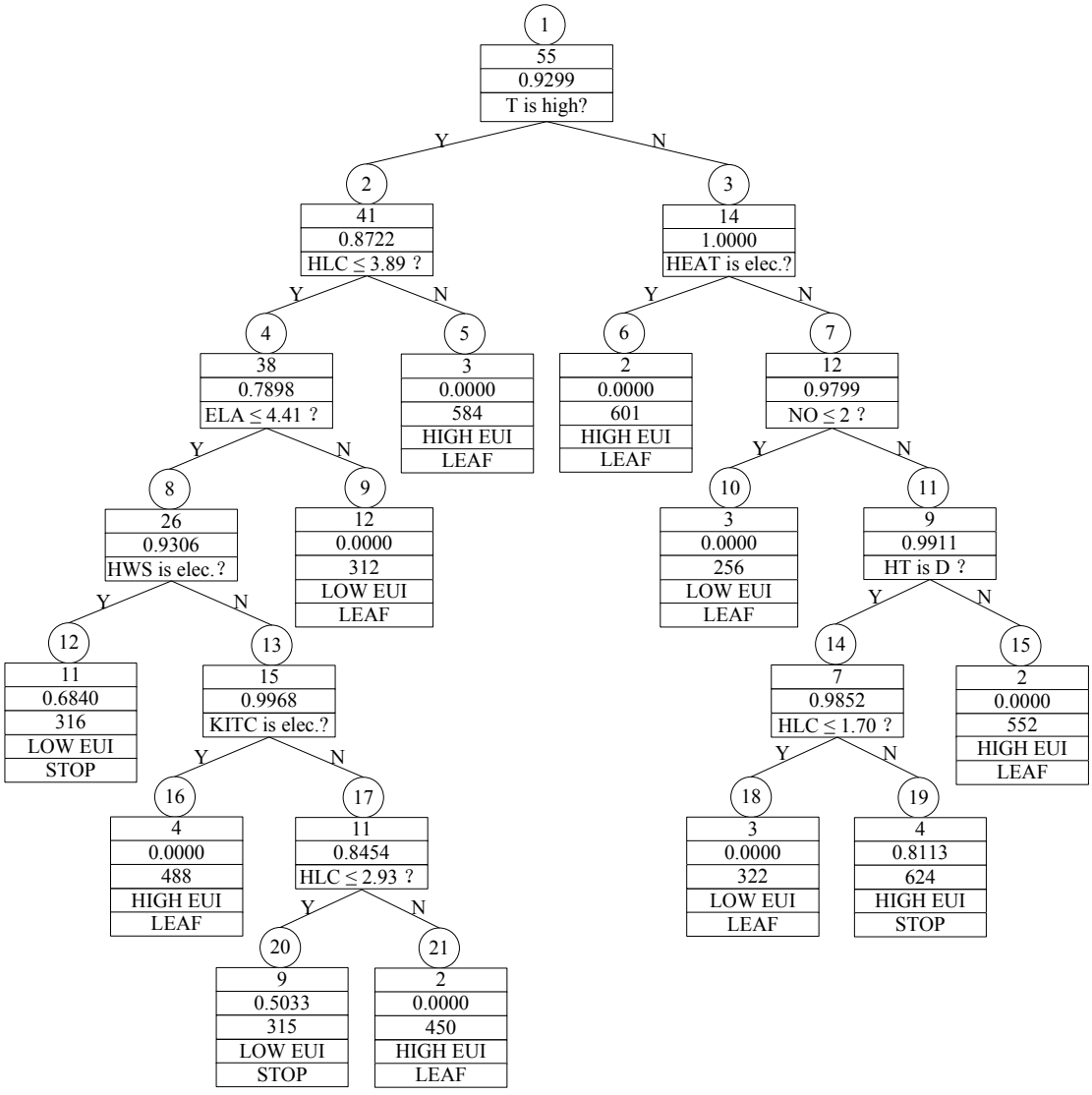

Figure 4-3 Decision tree for the prediction of building EUI level

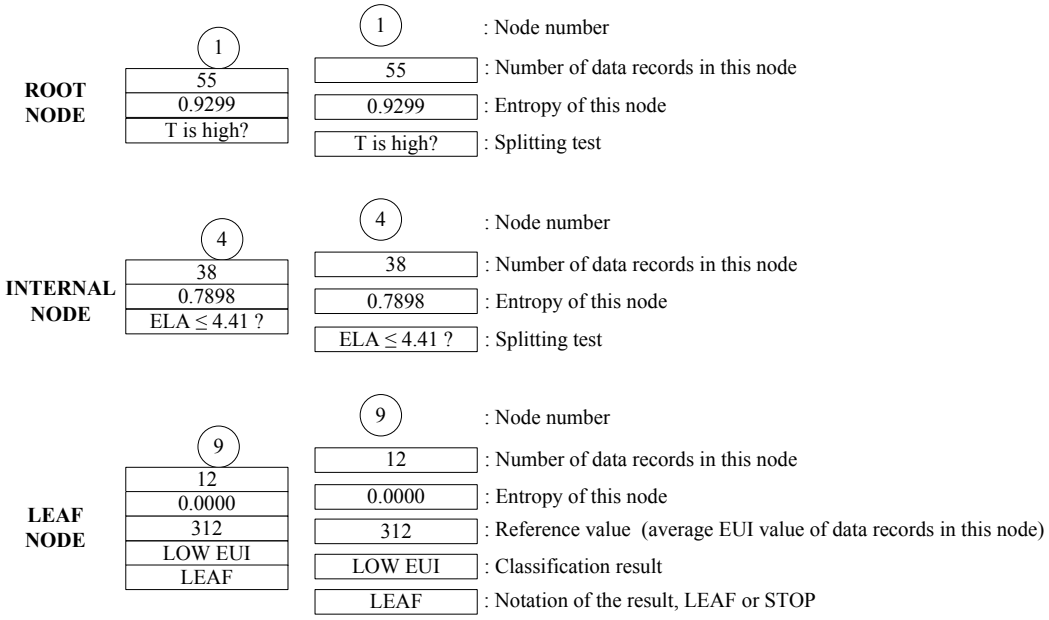

Figure 4-4 Explanatory note of decision tree nodes

Table 4-2 Decision rules derived from the obtained decision tree

	Node	Decision rules
1	5	If T is high and HLC > 3.89 then EUI is HIGH
2	6	If T is low and HEAT is electric then EUI is HIGH
3	9	If T is high and HLC ≤ 3.89 and ELA > 4.41 then EUI is LOW
4	10	If T is low and HEAT is non-electric and NO ≤ 2 then EUI is LOW
5	12	If T is high and HLC ≤ 3.89 and ELA ≤ 4.41 and HWS is electric then EUI is LOW
6	15	If T is low and HEAT is non-electric and NO > 2 and HT is apartment then EUI is HIGH
7	16	If T is high and HLC ≤ 3.89 and ELA ≤ 4.41 and HWS is non-electric and KITC is electric then EUI is HIGH
8	18	If T is low and HEAT is non-electric and NO > 2 and HT is detached and HLC ≤ 1.70 then EUI is LOW
9	19	If T is low and HEAT is non-electric and NO > 2 and HT is detached and HLC > 1.70 then EUI is HIGH
10	20	If T is high and HLC ≤ 3.89 and ELA ≤ 4.41 and HWS is non-electric and KITC is non-electric and HLC ≤ 2.93 then EUI is LOW
11	21	If T is high and HLC ≤ 3.89 and ELA ≤ 4.41 and HWS is non-electric and KITC is non-electric and HLC > 2.93 then EUI is HIGH

As mentioned previously, the decision tree accuracy should be evaluated to estimate how accurately it can predict building EUI levels before applying it to new residential buildings. Accordingly, the obtained decision tree was applied to the test dataset and the results are given in Table 4-3.

Table 4-3 shows that among twelve data records included in the testing set eleven records, accounting for 92%, are correctly classified. Given that the size of testing set is relatively small and only one record is misclassified, this accuracy is acceptable. At the same time, the WEKA analysis report also provides a confidence level for the classification of each data record. The confidence level determines how likely the test data record falls into that class and, it is equal to the ratio of the number of correctly classified data records to the total record number in that class in the training set. It can be seen from Table 4-3 that generally the confidence level for the classification is higher than 80%, indicating that most predictions are reliable. Furthermore, by using a pre-specified threshold, e.g., 80%, the confidence level could improve estimated accuracy of classification. In particular, if the confidence level of a data record classification exceeds the threshold, this classification is accepted. For example, if the threshold in this evaluation is set at 80%, then all the records, except the record 2 that is misclassified, will be accepted. Similarly, the threshold is very useful when applying decision rules to the prediction of new data sets. In addition, the error rate between the actual EUI value and the reference EUI value are also given in this table for the reliability test of reference value. It can be seen that, among 11 correctly classified

data records, five have an error rate lower than 5% while the other 6 have an error rate between 20% and 35%. This indicates that a higher concept hierarchy for building EUI needs to be formed to improve the prediction performance of reference value. However, this is limited by the size of database in this study.

Table 4-3 Results of decision tree accuracy evaluation

	Actual level	Predicted level	Correct or incorrect	Confidence level	Actual EUI	Reference EUI	Error
1	HIGH	HIGH	Correct	100%	449	450	0.2%
2	LOW	HIGH	Incorrect	75%	258	624	141.9%
3	HIGH	HIGH	Correct	100%	581	584	0.5%
4	LOW	LOW	Correct	100%	327	322	1.5%
5	HIGH	HIGH	Correct	100%	707	552	22.0%
6	LOW	LOW	Correct	81.80%	303	316	4.3%
7	LOW	LOW	Correct	81.80%	238	316	32.8%
8	LOW	LOW	Correct	88.90%	258	315	22.1%
9	HIGH	HIGH	Correct	100%	507	488	3.7%
10	HIGH	HIGH	Correct	100%	495	601	21.4%
11	LOW	LOW	Correct	81.80%	427	316	26.0%
12	HIGH	HIGH	Correct	100%	458	601	31.2%

4.3.3 Utilization of decision tree

Using decision tree for prediction

Based on predictor variables, decision tree and decision rules can be utilized to predict target variables. Assume the EUI level of a new residential building in Japan must be predicted by using the decision tree in Figure 4-3. The threshold of confidence level is set

at 85%. The typical building parameters are shown in Table 4-4. The building EUI level is predicted as follows:

Step 1 The root node, i.e., node 1 in this decision tree, is the starting point of prediction. From node 1, it can be seen the value of T should be first examined. Since T is high, the node 1 test *T is high* is satisfied, then go to node 2;

Step 2 Examine the value of HLC. Since HLC = 2, the node 2 test $HLC \leq 3.89$ is satisfied, then go to node 4;

Step 3 Examine the value of ELA. Since ELA = 3, the node 4 test $ELA \leq 4.41$ is satisfied, then go to node 8;

Step 4 Examine the value of HWS. Since HWS is non-elec., the node 8 test *HWS is elec.* is not satisfied, then go to node 13;

Step 5 Examine the value of KITC. Since KITC is gas, the node 13 test *KITC is elec.* is not satisfied, then go to node 17;

Step 6 Examine the value of HLC. Since HLC = 2, the node 17 test HLC ≤ 2.93 is satisfied, then go to node 20;

Step 7 Node 20 is a leaf node. As a result, the decision tree in Figure 4-3 predicts that the EUI level of the residential building is LOW. In this node, the correctly classified data records account for 89% and thus the confidence level of prediction is 89% that is larger than the predetermined threshold (85%). Therefore, the prediction is accepted. Furthermore, the value of correctly classified records in this node ranges from 242 MJ/m^2

to 389 MJ/m^2 and the average value is calculated at 315 MJ/m^2. These values can be used as reference values for the prediction, as mentioned previously.

Table 4-4 Building parameters for the prediction of building EUI levels

Number	Variable	Attribute value	Unit
1	T	High	
2	HT	Detached house	
3	CO	Wood	
4	NO	4	
5	FA	100	m^2
6	HLC	2	W/m^2K
7	ELA	3	cm^2/m^2
8	HEAT	Electricity	
9	HWS	Non-electricity	
10	KITC	Gas	

Model interpretation and knowledge extraction

Useful knowledge can be extracted from the decision tree based model so as to help understand energy consumption patterns and optimize a building design plan. For example, various parameters are automatically selected as predictor variables by the decision tree algorithm for the classification of EUI levels. These parameters are used to split the nodes of the decision tree and their degrees of closeness to the root node indicate the strength of the influence and the number of records impacted. Therefore, by examining the decision tree nodes, the significant factors, as well as their ranks, that determine the building energy demand profiles can be identified. In particular, the variable importance of this decision

tree model can be analyzed as follows: first, the root node, i.e., T, indicates that outside air temperature is the most important determinant of energy demand among all these factors. Then, for clarity, the significant factors for the high temperature districts (i.e. Hokuriku, Kanto, Kansai and Kyushu) and low temperature districts (i.e. Hokkaido and Tohoku) are identified separately and summarized in Table 4-5.

Table 4-5 Summary of significant factors

Potential factors	High temperature districts		Low temperature districts	
	Significant factors	Rank	Significant factors	Rank
House type			√	3
Number of occupants			√	2
Floor area				
Heat loss coefficient	√	1	√	4
Equivalent leakage area	√	2		
Construction type				
Space heating mode			√	1
Hot water supply mode	√	3		
Kitchen energy mode	√	4		

Clearly, four significant influencing factors are identified for each district and the only parameter found to be significant for the both districts is heat loss coefficient. This implies that the significance of these factors, except building heat loss coefficient, is dependent on outside air temperature. Moreover, among the three household appliance energy source parameters, space heating plays a role in low temperature districts while hot water supply and kitchen are significant in high temperature districts. Note that floor area and

construction types do not appear in the decision tree. This is reasonable since the target variable, i.e., EUI level, is a measure of annual total energy normalized for floor area and building heat loss coefficient embodies the effect of construction type. At the same time, these significant factors are ranked in terms of the degree of closeness to the root node. It can be found that the heat loss coefficient and space heating mode rank the first in the two districts respectively, and thus deserve extra attention when designing energy efficient buildings.

The decision tree provides the combinations of significant factors as well as the threshold values that will lead to high building energy performance. Based on such combination and threshold values, some hidden yet useful knowledge can also be extracted to help understand building energy consumption patterns. For example, it can be seen that, in high temperature districts, a higher building heat loss coefficient than 3.89 W/m^2K will normally cause a high EUI. Meanwhile, for a residential building with heat loss coefficient lower than 3.89 W/m^2K, a high equivalent leakage area (> 4.41 cm^2/m^2) will benefit energy conservation. This seems perhaps unreasonable and one possible explanation is that the high temperature districts locate in moderate climate and have a moderate outside air temperature range. Accordingly, in summer infiltration can serve as cooling source to remove the excess heat generated indoor, thereby reducing overall energy consumption. This indicates that a rational combination of heat loss coefficient and equivalent leakage area of residential buildings in high temperature districts is important to improve building

energy performance. Also, a further study on the range selection of equivalent leakage area may provide deeper insights into its impact on the building energy demand. Additionally, from the nodes 8 and 13 in Figure 4-3, it can be observed that the change of the energy source of hot water supply and kitchen will cause a substantial increase or decrease in the EUI. Clearly electrical water heaters, instead of non-electric water heaters such as natural gas heaters, should be used to save energy. Moreover, electrical water heaters can take full advantage of cheap nighttime electricity and thus help users save money spent on energy.

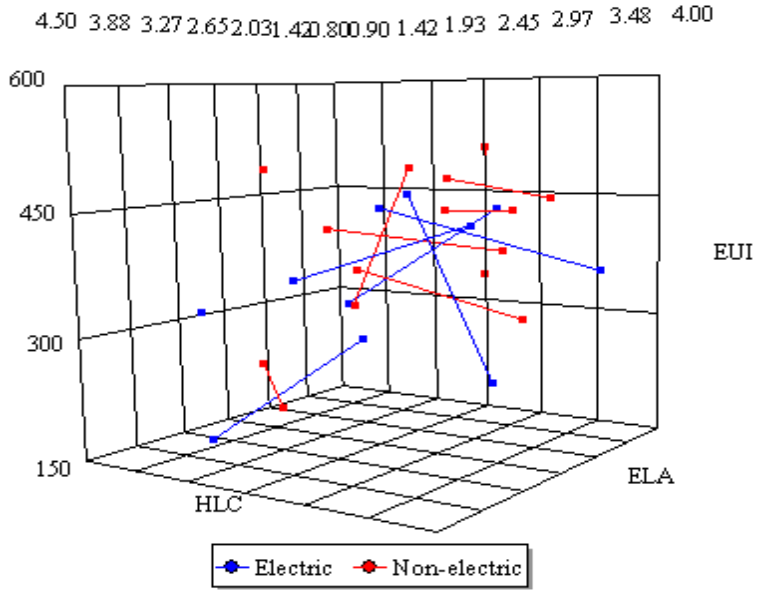

Figure 4-5 Comparison of the EUI between electric HWS and non-electric HWS

The EUI values in node 8 are plotted in Figure 4-5 in order to make a comparison between buildings with the electric HWS and buildings with the non-electric HWS. The

two significant factors with higher ranks than HWS, i.e., HLC and ELA, are also taken into consideration (HLC at abscissa, ELA at ordinate). The abscissa-ordinate plane is divided into various grids so that EUI values can be compared based on similar HLC and ELA values, thereby removing the impact of these two factors. It is apparent from Figure 4-5 that, in a same grid or adjacent grids, red points, which denote EUI values with non-electric HWS, are generally higher than blue points, which denote EUI values with electric HWS. This is in accordance with the above conclusion drawn from the decision tree.

With regard to kitchen energy source, electrical appliances, however, tend to consume more energy than the appliances using natural gas. This may have occurred since the power of many kitchen electrical appliances, such as rice cookers, is comparatively high and the use of these appliances is routine. Further, compared to hot water supply energy source, kitchen energy source has a smaller contribution to building energy demand and even though non-electric appliances are adopted in kitchen, an extra requirement on heat loss coefficient (≤ 2.93 W/m^{2}K) still need to be met in order to achieve low EUI levels.

In low temperature districts, from an energy saving point of view, building owners and designers should give a prior consideration to space heating energy source that plays a significant role in influencing EUI. The node 3 in Figure 4-3 shows that non-electric fuel, particularly kerosene and natural gas, should be used as primary source of residential space heating since the use of electric space heating tends to bring about a high EUI. This may be partly ascribed to the high efficiency of non-electric space heating devices such as

kerosene space heaters. Moreover, non-electric heating devices are more applicable than electric space heaters, such as air conditioners, in real life due to the high electricity rate in Japan. Similar to Figure 4-5, EUI values in node 3, together with EUI values in low temperature districts in the test dataset, are plotted in Figure 4-6. HLC and NO are used as abscissa and ordinate. The red and blue points represent EUI values with electric and non-electric space heating respectively. It can be observed that red points are generally higher than blue points. This observation is in accordance with above conclusion.

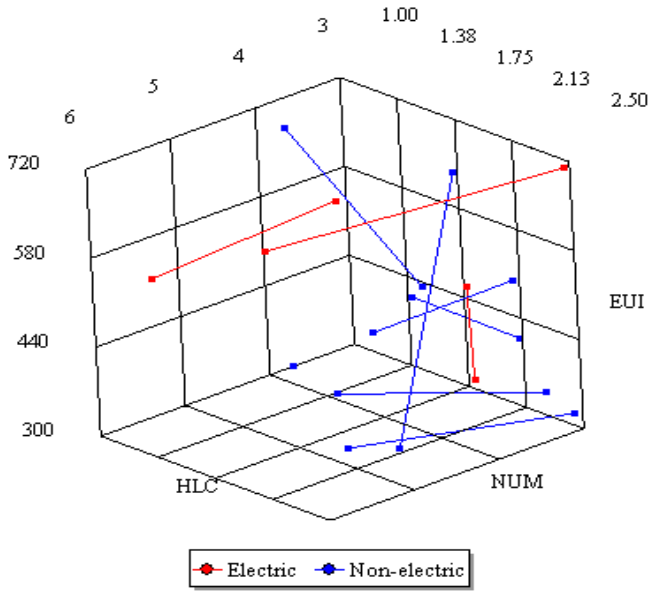

Figure 4-6 Comparison of EUI between electric HEAT and non-electric HEAT

The number of occupants, is another important determinant of EUI in low temperature districts. As can be seen, buildings with more than two occupants will have significantly

higher EUI than those with two occupants. This may have occurred since a larger family size will cause more complicated occupant behavior patterns thereby resulting in an increase in EUI. With regard to house type, detached houses with low heat loss coefficients ($\leq$ 1.70 W/m^{2}K) tend to have a better energy performance than apartments, which can occur for at least two reasons. First, a small HLC contributes greatly to reduce energy consumption on space heating and cooling; second, detached houses normally have larger areas than apartments while both of them have approximately same family size, which also lowers EUI values.

Such knowledge can help building designers and owners make intelligent decisions to improve building energy performance and reduce building energy consumption. For example, based on above knowledge, architects and building designers can identify the parameter that deserves more attention as well as its value range at the early design stage. Also, they can perform a fast performance estimation of newly constructed residential buildings. Moreover, building owners will easily determine which energy source should be used for space heating, hot water supply, and kitchen to save energy. It should be mentioned that heat loss coefficient and equivalent leakage area cannot be determined directly by architects and building designers. However, their value can be adjusted through some indirect measures such as improving construction material and building air tightness.

4.4 Summary

The decision tree method is applied to the Japanese residential buildings for predicting

and classifying building EUI levels and its basic steps, such as the generation of decision tree based on training data and the evaluation of decision tree based on test data are presented. The results have demonstrated that the use of decision tree method can classify and predict building energy demand levels accurately (93% for training data and 92% for test data), identify and rank significant factors of building EUI levels automatically, and provide the combination of significant factors as well as the threshold values that will lead to high building energy performance. The methodology along with derived knowledge could benefit building owners and designers greatly and one crucial benefit is improving building energy performance and reducing energy consumption and the money spent on energy. Although the decision tree method is mainly employed to predict categorical variables (the number of the predetermined target intervals depends on the size of database while too many intervals may result in errors in classification) and reference value (i.e., average value of EUI in each class in this study) instead of the precise value of target variables, as a modeling technique, the utilization of decision tree method is very simple and its result can be interpreted more easily compared to other widely used modeling techniques, such as regression methods and ANN methods.

5. A SYSTEMATIC PROCEDURE FOR STUDYING THE INFLUENCE OF OCCUPANT BEHAVIOR ON BUILDING ENERGY CONSUMPTION

5.1 Introduction

Efforts have been devoted to the identification of the impacts of occupant behavior on building energy consumption. However, various factors influence building energy consumption at the same time, leading to the lack of accuracy when identifying the individual effects of occupant behavior. As mentioned previously, one possible approach to estimating the effects of occupant behavior is to analyze measured building-related data. This study develops a methodology for identifying the effects of occupant behavior on building energy consumption through data analysis, thereby evaluating the energy saving potential by modifying user behavior and providing deep insights into the building energy consumption patterns.

5.2 Methodology

A methodology is proposed for examining the effects of occupant behavior on building energy consumption. Basically, it is realized by clustering similar buildings into groups based on the four influencing factors unrelated to user behavior (refer to Section 2.4), so that for each building in the same group the four factors have similar effects on

the building energy consumption. The effects of occupant behavior on building energy consumption can be identified accurately within each group. Furthermore, provided that there is a sufficient sample size and subject buildings have a large divergence in the four influencing factors, implying that the full effects of the four factors in each group can be similar enough and the energy consumption difference caused by them is comparatively small, energy consumption difference between buildings in each group could be thought of as being caused only by occupant behavior. The identification of building groups is the most important element of this methodology. Such identification is achieved mainly via cluster analysis, which was introduced in Section 3.3.2.

Before conducting cluster analysis, some preprocessing steps are needed to deal with the inconsistencies of different attributes. For example, most of the energy-related attributes have their own units. Switching attribute units from one to another may significantly change the attribute values, thereby impacting the quality and accuracy of clusters. Therefore, data transformation techniques are applied in order to help avoid dependence on the selection of attribute units. Also, data transformation can help prevent attributes with large ranges from outweighing those with comparatively smaller ranges. The contribution of different attributes to the building energy consumption may differ considerably. Thus, after data normalization, each attribute should be associated with a weight that reflects its significance. Grey relational analysis is used to identify such weights. The procedure of data transformation and grey relational analysis is introduced

in the following two sections.

5.2.1 Data transformation

As mentioned previously, data transformation was applied to deal with the inconsistencies in measured dataset. Specifically, min-max normalization (Han et al., 2006) is performed to scale the values to fall within a predetermined range. The main advantage of min-max normalization is its ability to reserve the relationships between the initial data since it carries out a linear normalization. Assume that x_{max} and x_{min} are the original maximum and minimum values of a numerical attribute. By min-max normalization, a value x of this attribute can be transformed to x' in the new specified range $[x'_{min}, x'_{max}]$ by calculating

$$x' = \frac{x - x_{min}}{x_{max} - x_{min}} (x'_{max} - x'_{min}) + x'_{min} \quad [5\text{-}1]$$

In this study, the new range is defined as [0, 1].

For binary attributes, their two states, such as the operation states of room air conditioners, i.e., [ON, OFF], can be transformed to [0, 1] or [1, 0] directly. The decision to recode these two states to either [0, 1] or [1, 0] depends upon whether or not there is a preferred positive value.

For multi-valued categorical attributes with an implicit order, it is often necessary to rank their ordered states first, and then map the obtained range onto [0, 1] by

$$x_i' = \frac{rank_i - 1}{rank_{max} - 1} \quad [5\text{-}2]$$

where

x_i': transformed value of each state

$rank_i$: corresponding rank of each state

$rank_{max}$: maximum rank

For example, the four levels of certification in the Leadership in Energy and Environmental Design (LEED) Green Building Rating System, i.e., [CERTIFIED, SILVER, GOLD, PLATINUM], are transformed to [0, 1/3, 2/3, 1] using the above method.

5.2.2 Grey relational analysis

Grey relational analysis (GRA) was proposed to find grey relational grades and a grey relational order (i.e., the rank of grey relational grades) that can be used to describe primary trend relationships between related factors, and to identify the important factors that significantly influence predefined target factors (Deng, 1989). For example, if the building energy consumption is defined as the target factor, GRA can provide grey relational grades for its various influencing factors, such as outdoor air temperature and floor area. These grey relational grades are numerical measures of the impact of the influencing factors on total building energy consumption. Larger grey relational grades indicate more significant impacts. The main advantages of GRA over other similar multi-factorial analysis methods such as regression analysis and principal component

analysis are its comparative simplicity and the ability to deal with small data sets that do not have typical probability distributions.

Let y_0 be the objective sequence (measured data of target factors, such as building energy consumption) and y_i be the compared sequences (measured data of related factors, such as the various influencing factors of building energy consumption):

$$y_0 = \left(y_0(1), y_0(2), \dots, y_0(n)\right) \quad [5\text{-}3]$$

$$y_i = \left(y_i(1), y_i(2), \dots, y_i(n)\right),\ i = 1, 2, \dots, m \quad [5\text{-}4]$$

The procedure of GRA is described as follows:

Step 1 Normalize the raw data (Min-max normalization is used in this study), y_0 and y_i are used to denote obtained normalized sequences;

Step 2 Calculate grey relational coefficients ξ. $\xi_i(k)$ between y_0 and y_i is defined as:

$$\xi_i(k) = \frac{\min_i \min_k |y_0(k) - y_i(k)| + \alpha \max_i \max_k |y_0(k) - y_i(k)|}{|y_0(k) - y_i(k)| + \alpha \max_i \max_k |y_0(k) - y_i(k)|} \quad [5\text{-}5]$$

$$i = 1, 2, \dots, m;\ k = 1, 2, \dots, n$$

where α is the distinguishing coefficient and $0<\alpha<1$, normally $\alpha = 0.5$;

Step 3 Calculate the grey relational grade γ:

$$\gamma(y_0, y_i) = \frac{1}{n}\sum_{k=1}^{n} \xi_i(k) \quad [5\text{-}6]$$

Step 4 Rank the obtained grey relational grades; thus, grey relational order can be identified.

As mentioned previously, grey relational grade is employed to be weighted

coefficients of the corresponding attributes in cluster analysis. Note that grey relational grades range from 0 to 1. Generally, $r > 0.9$ indicates a marked influence, $r > 0.8$ indicates a relatively marked influence, $r > 0.7$ indicates a noticeable influence, and $r < 0.6$ indicates a negligible influence (Fu et al., 2001).

5.3 Selection of typical parameters

Table 5-1 Representative parameters of the four influencing factors

Influencing factors	Representative parameters	Category	Unit	Abbreviation
City Climate	(i) Annual mean air temperature	numerical	°C	T
	(ii) Annual mean relative humidity	numerical		RH
	(iii) Annual mean wind speed	numerical	m/s	V
	(iv) Annual mean global solar radiation	numerical	MJ/m^2	RA
Building-related characteristics	(i) House types[a*]	categorical		HT
	(ii) Building area	numerical	m^2	FA
	(iii) Equivalent leakage area[b*]	numerical	cm^2/m^2	ELA
	(iv) Heat loss coefficient[c*]	numerical	W/m^3K	HLC
User-related characteristics except social and economic factors	(i) Number of occupants	numerical		NO
Building services systems and operation[d*]	Energy source of usage for			
	(i) Space heating and cooling	categorical		HEAT
	(ii) Hot water supply	categorical		HWS
	(iii) Kitchen equipment	categorical		KITC

a) House types are divided into either detached house or apartment.*

b) Measured by the fan pressurization method.*

c) Calculated based on building design plans.*

d) Energy source of usage is divided into either electric or non-electric. Since all of the space cooling equipment is electric, the value of HEAT is determined by space heating equipment.*

The applicability of the proposed methodology is demonstrated by the measured data collected from the Japanese residential buildings. The main parameters that could

generally represent the four influencing factors unrelated to the occupant behavior should be identified before the cluster analysis. Based on the characteristics of the residential buildings in Japan, twelve representative parameters of the four influencing factors were captured from the database and are outlined in Table 5-1.

5.4 Results and discussion

5.4.1 Grey relational grades

The goal of this research is to identify the influences of the occupant behavior on the building energy consumption. Therefore, annual building energy use intensity (EUI) in 2003 was selected as the objective sequence in GRA, and accordingly, there is no need to consider the building area independently. Among the remaining eleven parameters, four weather parameters are time-series variables that can be viewed as a function of time. In order to take both the impact of season and regional climate difference into consideration, grey relational grades were first calculated for each building based on monthly building EUI and local monthly weather parameters (Japan Meteorological Agency 2003); then, an average was taken over grey relational grades in each district. For the other seven parameters, considering the size of database, grey relational grades were calculated on all the buildings.

The results of GRA are given in Table 5-2. It shows that, generally outdoor air temperature influenced EUI more significantly than the other three parameters, especially

in the cold districts, i.e., Hokkaido and Tohoku. Also, the number of occupants and the heat loss coefficient had noticeable impact on the building energy performance, since the grey relational grades of these two parameters are between 0.7 and 0.8. This implies that these two parameters deserve more attention in the building design phase.

Table 5-2 Grey relational grades for each district

District	Grey relational grades										
	T	V	RH	RA	NO	HLC	ELA	HT[a*]	HEAT[b*]	HWS[b*]	KITC[b*]
Hokkaido	0.799	0.584	0.620	0.683	0.701	0.780	0.490	0.617	0.537	0.514	0.551
Tohoku	0.831	0.555	0.765	0.662							
Hokuriku	0.772	0.532	0.644	0.716							
Kanto	0.737	0.601	0.732	0.641							
Kansai	0.712	0.580	0.695	0.690							
Kyushu	0.654	0.605	0.661	0.675							

a[*] The two states of house types, i.e., detached house and apartment, are transformed to [0, 1].

b[*] The two states of these three parameters, i.e., electrical and non-electrical, are transformed to [0, 1].

5.4.2 Cluster analysis

After data preprocessing and the calculation of the grey relational grades, i.e., weighted coefficients of the selected parameters in Table 5-2, cluster analysis was conducted using the open-source data mining software WEKA. The results of cluster analysis are given in Table 5-3. With the consideration of the size of the database, four clusters were determined by the *K*-means algorithms based on Euclidean distance measures. Cluster centroids, representing the mean value for each dimension, were used to characterize the clusters.

Table 5-3 Centroid of each cluster and statistics on the instances in each cluster

Attribute	Full Data	Cluster			
		1	2	3	4
T	0.451	0.609	0.483	0.312	0.408
V	0.313	0.316	0.303	0.339	0.302
RH	0.395	0.262	0.417	0.428	0.439
RA	0.347	0.318	0.370	0.343	0.343
HT	0.166	0.000	0.134	0.411	0.116
HLC	0.183	0.254	0.154	0.116	0.229
ELA	0.394	0.291	0.413	0.460	0.390
NO	0.275	0.216	0.320	0.234	0.296
HEAT	0.305	0.331	0.000	0.501	0.537
HWS	0.307	0.514	0.067	0.514	0.289
KITC	0.222	0.551	0.000	0.514	0.000
Clustered instances and proportion	67 (100%)	13 (19%)	23 (34%)	15 (22%)	16 (24%)

For example, cluster 1, in comparison with the other clusters, is a segment of buildings representing a high outdoor air temperature (the cluster centroid of T in this cluster is 0.609, which is higher than that in the other three clusters), detached houses (the cluster centroid of HT in this cluster is 0, indicating that all the buildings in this cluster are detached house), high heat loss coefficients, low equivalent leakage areas, small number of occupants, non-electrical hot water supplies and kitchen equipment, etc. Similarly, the other clusters can be explained as follows: cluster 2 is characterized as high solar radiation, large number of occupants, electrical space heating and cooling, and electrical kitchen equipment. Cluster 3 is a segment of buildings representing a low outdoor air temperature, low heat loss coefficients, high equivalent leakage area, and non-electrical hot water supplies. Cluster 4 is characterized as high outdoor relative

humidity, non-electrical space heating and cooling, and electrical kitchen equipment. In addition, the centroid of all the data is also given for comparison with the cluster centroids, as shown in Full Data column in Table 5-3. The internal cohesion and external separation for the clusters based upon the eleven attributes imply that these attributes have the most similar holistic effects on the building energy performance in the same cluster, while the effects are significantly distinct for the buildings in different clusters.

5.4.3 Effects of occupant behavior

End-use load shapes

After the generation of four clusters, the end-use loads of the buildings in each cluster were averaged over one year. Figure 5-1 shows the average annual EUI of different end-use loads for each cluster. The proportion of each end-use load to the whole load is also given above the corresponding bar.

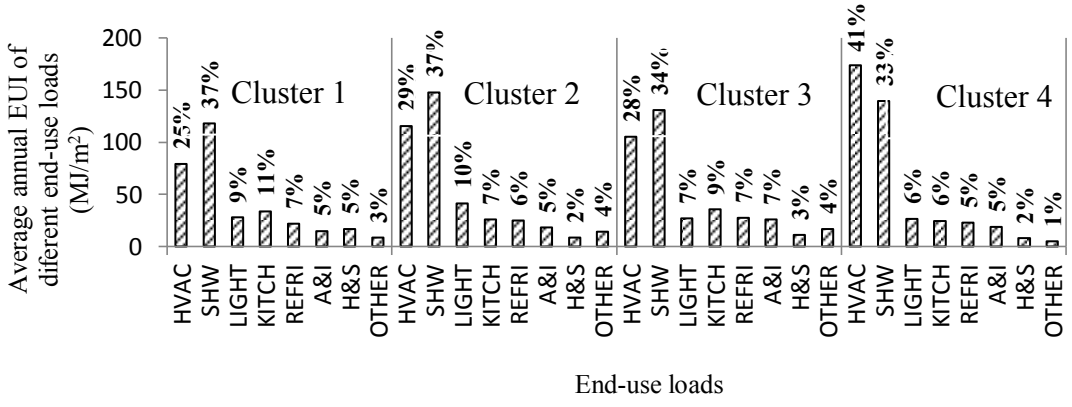

Figure 5-1 Average annual EUI of different end-use loads

Figure 5-1 shows that hot water supply and HVAC form the two largest categories of end-use loads in terms of average annual EUI in all four clusters, while housework and sanitary and 'others' have a modest contribution. Also, the two largest loads far exceed the other six end-use loads that do not have significant variations in the proportion among most of the clusters. This indicates that occupants in different clusters had similar behavior. Moreover, the proportions of both hot water supply and HVAC remain approximately steady among these clusters, except that there is a noticeable increase in the HVAC proportion in Cluster 4, which is mainly characterized by the medium-low outdoor air temperature and non-electrical space heating equipment. This increase may be partly caused by two factors: 1) the high electricity rate in Japan, and 2) the high energy efficiency of non-electrical space heating devices such as kerosene space heaters. A high electricity rate tends to restrict occupants' usage of electrical heating/cooling equipment in the other three clusters, while high efficiency of non-electrical space heating devices encourages occupants' utilization of them in Cluster 4, thereby increasing energy consumption. Therefore, a rational combination of electricity rates and primary heating/cooling sources could help reduce building energy consumption through influencing occupant behavior.

Variability in annual EUI of different end-use loads induced by occupant behavior

In order to examine the variability in the annual EUI of different end-use loads caused by occupant behavior, the end-use loads in each cluster were normalized and plotted.

Figure 5-2 depicts a box plot of normalized annual EUI of different end-use loads. The annual EUI of each building is normalized by the mean value of all the buildings in that cluster, thus highlighting the variability and allowing all the end-use loads to be plotted together on the same scale. As shown in Figure 5-2, a large variability that ranges from close to zero to about four times over the mean value is induced by the user behavior. Since the end-use loads in each building are normalized by the mean value of all the buildings in that cluster, the value of end-use loads ranges from zero to twice as many as the mean value was considered to be an insignificant variation. Accordingly, the threshold value for significant variation is defined as 2 (illustrated by the dash line). Except for SHW and REFRI, the range of the other six end-use loads exceeds the threshold value in most of the clusters. Such high variability implies that there still remains great potential for energy saving by improving occupant behavior related to these six domestic end-use loads. Contrarily, considering the relatively narrow range of SHW and REFRI, there could be little expectation of reducing energy consumption in these areas via improving occupant behavior.

Reference building and energy-saving potential

In order to evaluate the energy-saving potential for the four clusters, a reference building for each cluster was first defined. The characterization of each reference building was carried out by identifying the building with the energy consumption closest to the cluster energy consumption centroid in terms of Euclidean distance and end-use

loads. The annual EUI of different end-use loads of a reference building for each cluster is given in Table 5-4.

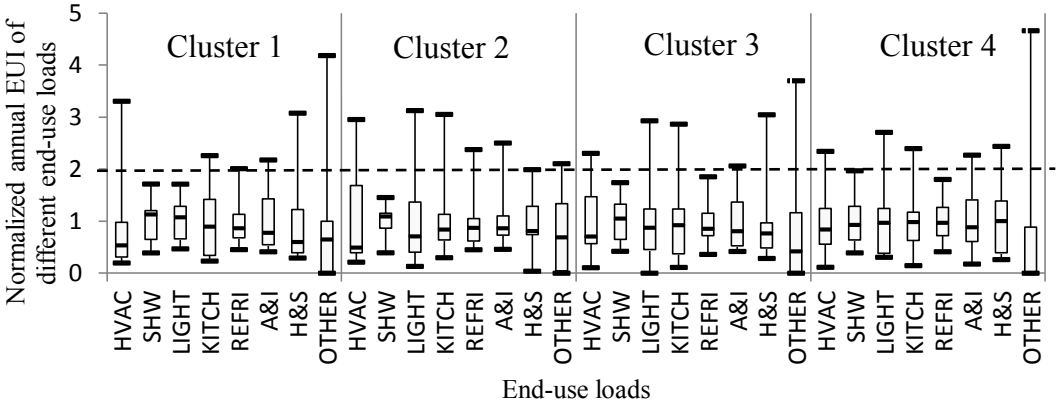

Figure 5-2 Boxplot of normalized annual EUI of different end-use loads

Table 5-4 Annual EUI of end-use loads of reference buildings (MJ/m^2)

	HVAC	SHW	LIGHT	KITCH	REFRI	A&I	H&S	OTHER	SUM
Cluster 1	77	165	31	24	25	12	29	0	363
Cluster 2	45	161	39	25	22	20	7	12	332
Cluster 3	154	141	33	42	20	13	6	0	409
Cluster 4	188	212	34	25	15	19	11	0	504

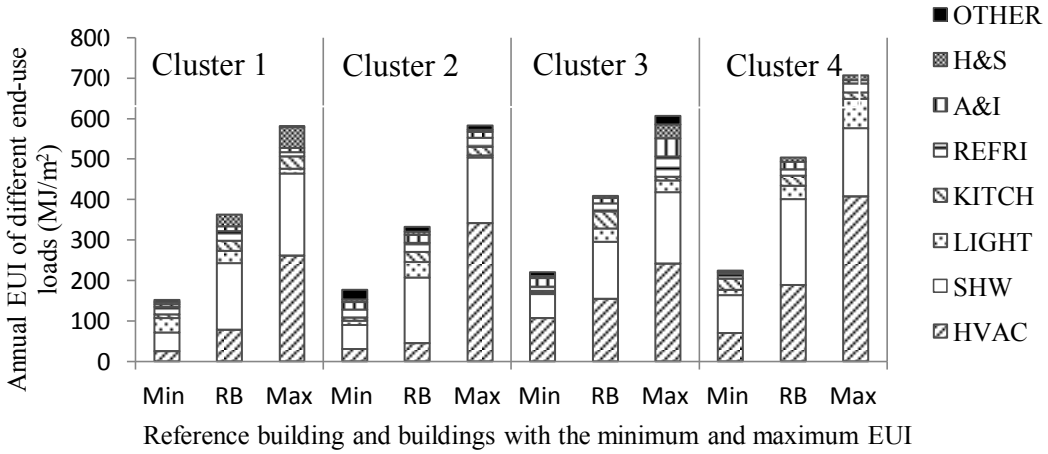

Figure 5-3 Stacked-column diagram of annual EUI of end-use loads of three typical buildings

Figure 5-3 shows the stacked-column diagram of annual EUI of different end-use loads of three typical buildings in the four clusters: a reference building (RB) and buildings with the minimum (Min) and maximum (Max) annual EUI. Occupant behavior led to a huge difference between these three different buildings in each cluster. In this study, annual EUI of different end-use loads of a reference building was taken as a baseline. Accordingly, the energy-saving potential of a building with a larger annual EUI than that of a reference building could be determined by computing the difference between them. For example, the potential energy savings that could be achieved by improving occupant behavior for the buildings with the maximum annual EUI in the four clusters, i.e., $EUI_{Max} - EUI_{RB}$, were 281 MJ/m^2, 250 MJ/m^2, 198 MJ/m^2, and 202 MJ/m^2, respectively. Moreover, comparison with a reference building provided a means of examining which end-use load seemed to have the greatest potential for energy conservation. For instance, comparison between the building with the maximum annual EUI and the reference building in each cluster indicated that HVAC contributed the most towards energy saving, while HWS had a negligible contribution. This result is consistent with the conclusion drawn from Figure 5-1. Similarly, other end-uses loads with noticeable energy-saving potential in each cluster could be identified, such as housework and sanitary in Cluster 1 and lighting in Cluster 4. Such information can help building owners realize which occupant behavior should be modified to improve building energy performance. Furthermore, based on this information, a better outcome may be achieved

if building occupants receive an energy-saving education and tips on how to improve their behavior. It should be noted that, in comparison with a reference building, buildings with the minimum annual EUI in the four clusters not only had lower HVAC EUI, but also had much smaller SHW EUI. A possible explanation for this is that occupants in these buildings reduced energy consumption by being concerned about the cost in living standards. For example, these occupants may decrease the frequency of utilization of room air conditioners in the cooling season, even though the indoor temperature is not the best comfort temperature. Further field investigation is needed to identify the real reasons.

Monthly variations of end-use loads induced by occupant behavior

In order to examine the effects of occupant behavior on end-use loads over time and buildings, monthly variations of average end-use loads in each cluster were plotted in semi-logarithmical graphs, as shown in Figures 5-4 to 5-7. Clearly HVAC shows a significant variation in all the four clusters. Generally, the peak of HVAC occurred in the heating season, especially in December and January, while the trough of HVAC occurred in the cooling season, especially June and July. This may have occurred because four districts (i.e., Hokuriku, Kanto, Kansai, and Kyushu) have a moderate climate and the other two (Hokkaido, Tohoku) are located in a cold climate, and cooling energy demand is considerably lower than heating energy demand. At the same time, HVAC in Cluster 3, characterized by the lowest outdoor air temperature, had the biggest peak-to-trough ratio.

This indicates that weather conditions significantly influenced occupant behavior, thereby impacting building energy consumption. With respect to SHW, its variation is noticeable, considering the absolute magnitude of the variation is comparatively large. In general, the peak of SHW occurred in December or January, while the trough occurred in August or September. Evidently this was also caused by weather conditions, especially outdoor air temperature. With regard to LIGHT, KITCH, REFRI, and A&I, these four curves bear a remarkable similarity to each other in the four clusters, and almost all of them vary by less than 20% from the mean. This indicates that these households tended to maintain their lifestyles, and the level of their general indoor activities associated with these end-use loads did not fluctuate wildly from month to month. In addition, the remaining two smaller end–use loads, i.e., H&S and OTHER, showed a marked seasonal variation in the four clusters, while the absolute magnitude of the variation is comparatively small. Basically the end-use loads in a heating season are higher than in a cooling season. A further investigation of corresponding occupant-behavior patterns is needed to explain the reasons for this variation.

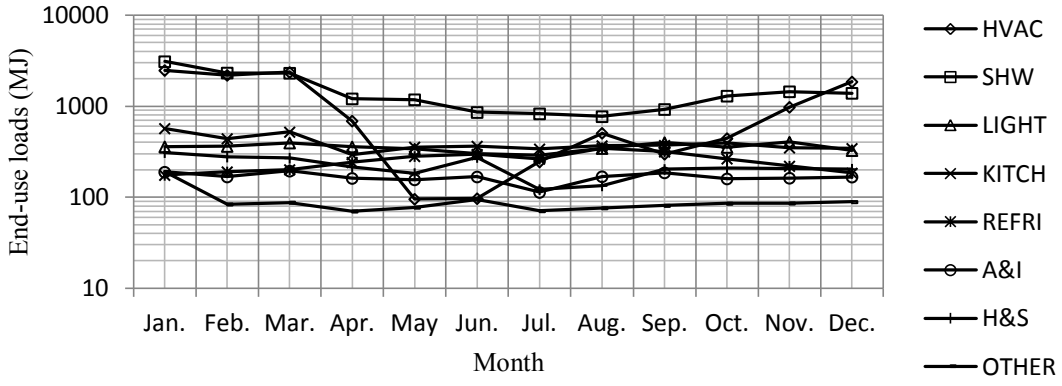

Figure 5-4 Monthly variation of end-use loads in Cluster 1

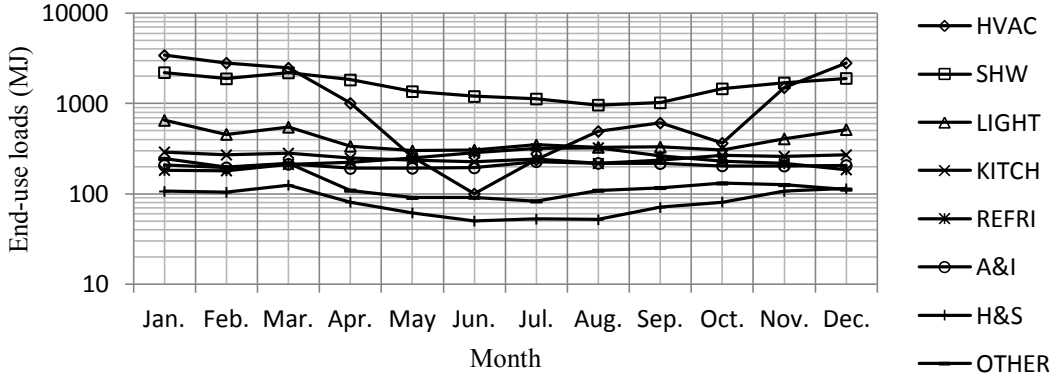

Figure 5-5 Monthly variation of end-use loads in Cluster 2

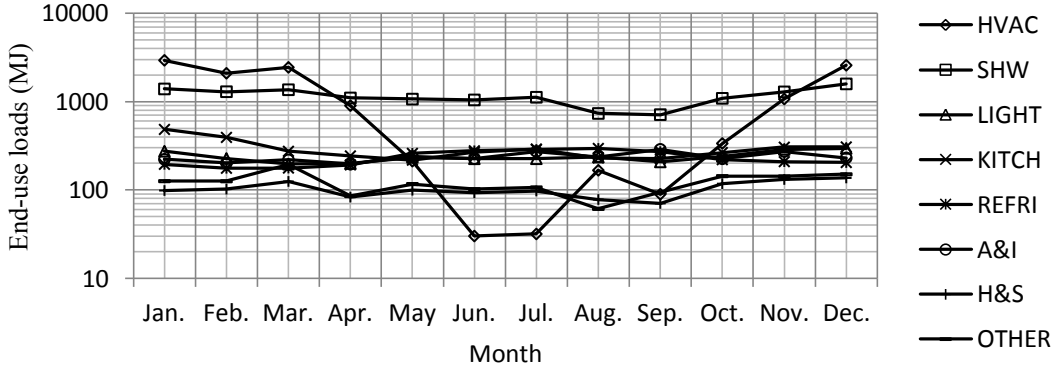

Figure 5-6 Monthly variation of end-use loads in Cluster 3

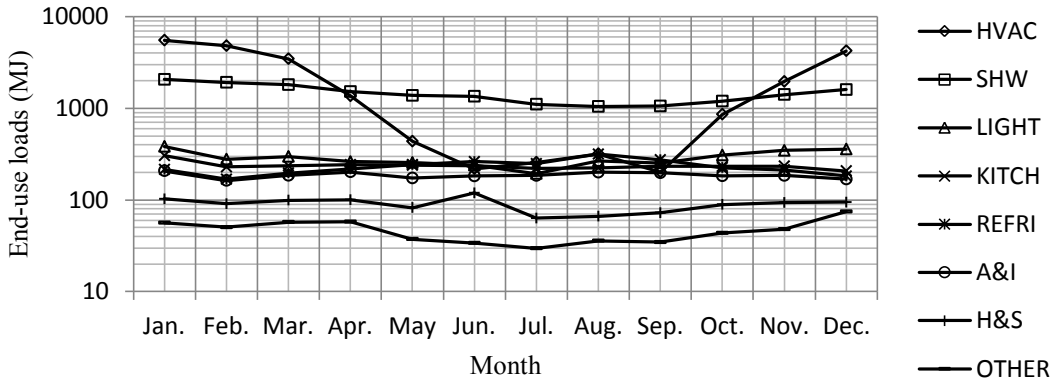

Figure 5-7 Monthly variation of end-use loads in Cluster 4

Monthly average indoor temperature of air-conditioned room

Different occupant behavior, especially those associated with HVAC, can significantly affect indoor climate, which in turn will have an influence on occupant behavior, thereby causing dramatic differences in building energy consumption. Therefore, the effects of occupant behavior on the building energy consumption should be understood and interpreted in relation with the investigation of indoor climate. Figures 5-8 to 5-11 show the monthly average living-room temperature of three typical buildings in each cluster: the reference building (RB) and buildings with the maximum and minimum annual EUI (Max and Min). These selected living rooms had air conditioners and/or heating equipment. As shown in Figure 5-8, there is a significant difference between living-room temperatures of the three buildings in the cooling season and a minor difference in other seasons. The living room of Max was maintained at a temperature of about 24 °C in the cooling season. At the same time, the room temperature

of Min was around 5 °C higher than that of Max, and the room temperature of RB was generally between that of Max and Min in this season. Considering that Cluster 1 is characterized by the highest outdoor air temperature, it can be deduced that the frequency of utilization of room air conditioners in the cooling season in these three buildings can be ranked as: Max > RB > Min. With respect to the other three clusters, Figures 5-9 to 5-11 show that the living room of Max was maintained at a temperature of about 24 °C throughout the year, while living-room temperatures of RH and Min varied with the outdoor air temperature. Clearly the frequency of utilization of space cooling/heating equipment in the three buildings in these three clusters has the same order as that in Cluster 1 for both heating and cooling seasons. These results suggest that occupant behavior that seeks thermal comfort normally results in high energy consumption. Therefore, there has to be a trade-off between human thermal comfort and building energy consumption, and it is necessary to strike a balance between achieving a high comfort level and reducing energy consumption through modifying occupant behavior.

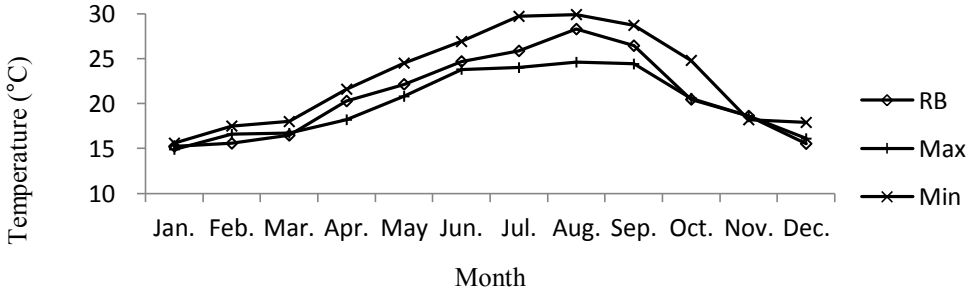

Figure 5-8 Monthly average living-room temperature of three typical buildings in Cluster 1

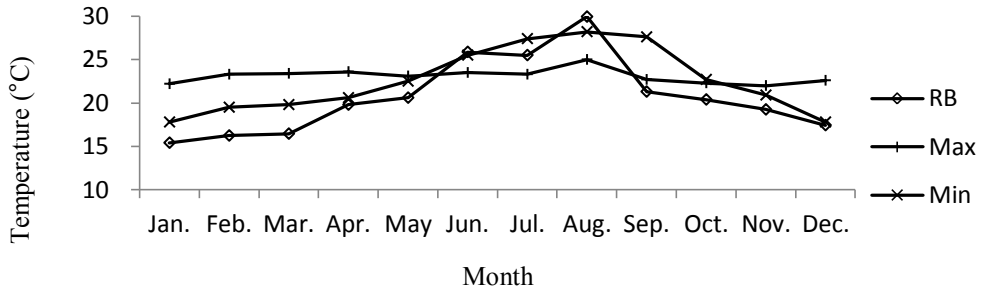

Figure 5-9 Monthly average living-room temperature of three typical buildings in Cluster 2

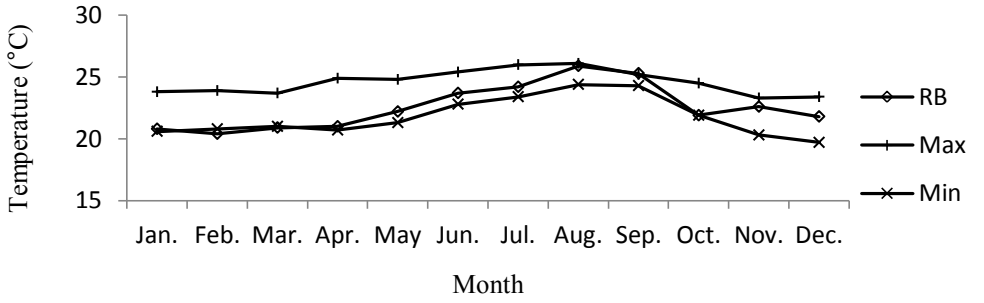

Figure 5-10 Monthly average living-room temperature of three typical buildings in Cluster 3

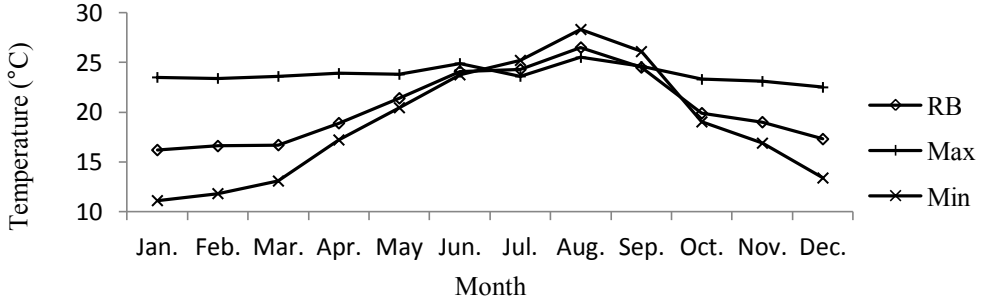

Figure 5-11 Monthly average living-room temperature of three typical buildings in Cluster 4

5. 4 Summary

This chapter presents the development of a novel data analysis methodology through clustering techniques for identifying the effects of occupant behavior on the building energy consumption. It is realized by organizing similar buildings among all the investigated buildings into various groups based on the four influencing factors unrelated to user behavior, so that for each building in the same group the four factors have similar full effects on energy consumption. Min-max normalization techniques are performed as a data preprocessing step to deal with the inconsistencies of different attributes. Grey relational analysis is also carried out, and grey relational grades, a measure of relevancy between two factors, are used as weighted coefficients of attributes in cluster analysis.

In order to demonstrate its applicability, this methodology was applied to the residential buildings located in six different districts of Japan. Twelve attributes were captured from the database to represent the influencing factors unrelated to occupant behavior. The *K*-means method was selected in cluster analysis and four clusters were obtained as a result.

In these four clusters, the effects of occupant behavior on the building energy consumption were examined at the end-use level. End-use variations over time and buildings induced by the occupant behavior were analyzed. Also, as a preliminary step toward identifying energy-saving potential, a reference building was defined as the building whose energy consumption was the closest to cluster energy consumption

centroid in terms of Euclidean distance and end-use loads. Moreover, indoor climate was investigated to better understand and interpret the effects of occupant behavior.

The proposed method allows researchers to evaluate building energy-saving potential by improving user behavior, and provides multifaceted insights into building energy end-use patterns associated with occupant behavior. The results obtained could help prioritize efforts of modification of occupant behavior to reduce building energy consumption.

6. A NOVEL METHODOLOGY FOR KNOWLEDGE DISCOVERY THROUGH MINING ASSOCIATIONS BETWEEN BUILDING OPERATIONAL DATA

6.1 Introduction

Building industry is not only energy-intensive, but also knowledge-intensive. Hence, it is highly desirable that useful knowledge hidden in building operation data be discovered to help reduce building energy consumption. In particular, associations and correlations between building operational data should be examined.

This chapter reports the development of a methodology for examining the associations and correlations between building operational data. The goal of this study is to achieve a better understanding of building operation and to provide opportunities for developing strategies to reduce energy consumption while maintaining a comfortable and healthy indoor environment.

6.2 Methodology

A methodology is proposed for examining all the associations and correlations between building operational data and leading to knowledge discovery. The methodology is based on a basic data mining technique: association rule mining (ARM), which was introduced in Chapter 3. In order to find and take advantage of more complete

associations and correlations, building operational data in two different time periods (i.e., both a day and a year) need to be mined separately, considering associations/correlations between operational data in different time periods could be significantly different. Moreover, data in two different years also need to be mined separately, and obtained associations/correlations in the two years should be compared between each other. The comparison can assist in identifying marked changes in associations/correlations and also building operation, thereby uncovering useful knowledge. The proposed methodology is given in Figure 6-1, and it can be divided into 8 steps and is explained as follows:

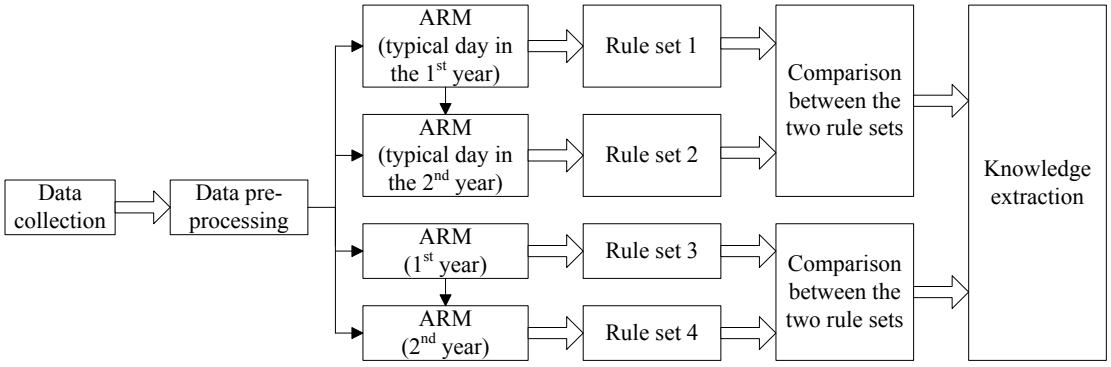

Figure 6-1 Proposed methodology to examine all the associations and correlations between building operational data

Step 1 Data collection. Two-year building operational data need to be collected and stored in a database.

Step 2 Data pre-processing. Measured data is often noisy (especially containing outlier values whose values are grossly different, i.e., much higher or lower, from others in databases), which can lead to low-quality mining results. Hence, the collected data should be processed to remove outliers.

Step 3 Perform the ARM in typical day (e.g., the coldest or hottest day) data in the 1st year. Obtained rules are stored in *Rule set 1* (see Figure 6-1).

Step 4 Select parameters having associations in the typical day data in the 1st year; and perform the ARM in the typical day data in the 2nd year within the selected parameters, in order to remove time effects and reduce other influences, such as the change of occupant behavior and weather conditions. Obtained rules are stored in *Rule set 2*.

Step 5 Perform the ARM in the 1st year data. Obtained rules are stored in *Rule set 3*.

Step 6 Select parameters having associations in the 1st year data; and perform the ARM in the 2nd year data within the selected parameters. Obtained rules are stored in *Rule set 4*.

Step 7 Compare the rules between the rule sets 1 and 2, and the rule sets 3 and 4; and highlight the similarity and difference in associations between the two different time periods (i.e., the typical day in the1st year and 2nd year, the 1st year and the 2nd year).

Step 8 Extract useful knowledge from the comparison between these rules.

For demonstration purposes, the proposed methodology was applied to the collected data in the EV building in this study.

6.3 Data pre-processing

Outliers are data objects in the database whose values are grossly different (i.e., much higher or lower) from others. Outliers regularly occur in building energy consumption measurement and they are often indicative of measurement errors, and thus must be removed. Removal of outliers plays a crucial role in preparing for the ARM, since the outliers will skew and thus alter the grouping of data. For example, suppose an attribute ranges from 0 to 10, and can be discretized into two intervals, [0, 5) and [5, 10] (or LOW and HIGH) as mentioned previously. If there exists an outlier (e.g., 30), then the two intervals are [0, 15) and [15, 30] (or LOW and HIGH) by using the same method. Accordingly, all the data are defined as LOW except the outlier, which is not true.

Various methods can be used for effective detection and removal of the outliers. In this study, a method based on the lower quartile (Q_1) and the upper quartile (Q_3) of the standard boxplot was used due to its simplicity (Han et al., 2006). Specifically, outlying values can be distinguished using the following two rules:

Rule 1: data values that are less than $Q_1 - 1.5 \times (Q_3 - Q_1)$ are defined as outliers

Rule 2: data values that are larger than $Q_3 + 1.5 \times (Q_3 - Q_1)$ are defined as outliers

Additionally, in order to perform the ARM, the value of quantitative attributes generally needs to be classified into categorical values. Given that building operational data, such as supply air temperature and monthly energy consumption, is normally described as either high or low by occupants in practice, a two-interval scale, i.e., HIGH

and LOW, was applied in this study. Specifically, for each quantitative attribute, data ranged from the average of the maximum and minimum to the maximum value is 'HIGH', and data ranged from the minimum value to the average of the maximum and minimum is 'LOW'.

With consideration of the seasonality of building energy consumption, the ARM was performed based on seasonal data instead of annual data in this study (refer to *steps* 5 and *6* in Section 6.2). Given that the EV building is located in Montreal which has very cold winters, the winter data in both 2007 and 2009 was mined to generate association rules to provide opportunities for saving more energy (as mentioned earlier, the winter data in 2008 was unavailable). Furthermore, only the data in working days/hours was used when mining seasonal data, considering that building energy consumption is significantly different between working days/hours and non-working days/hours due to occupant behavior (for the EV building, non-working days include weekends and holidays; and working hours are from 8 AM to 5 PM). The resulting data in 2007 and 2009 were stored in dataset_1 and dataset_2, respectively. Figure 6-2 shows the distribution of two intervals of the entire ARM attributes in the dataset_1 after the removal of outliers and discretization. Note that the numbers in the abscissa represent the ARM attributes, and correspond to the numbers in Table 1. Clearly, it can be observed that most of the percentages range from 30% to 70%, indicating a roughly uniform distribution.

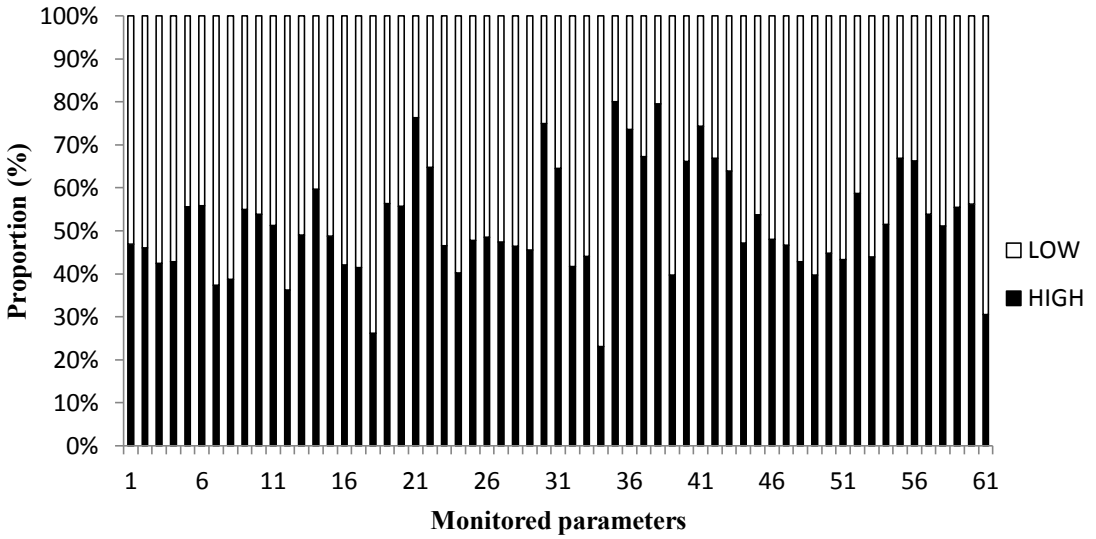

Figure 6-2 Distribution of two intervals of all monitored parameters in the dataset_1

6.4 Results and Discussion

6.4.1 ARM on the Coldest Day in the Dataset_1 and Dataset_2

The initial rule mining was carried out with the dataset_1 and dataset_2 on the coldest day in both 2007 and 2009. After experimenting with various combinations of *support* and *confidence* values, a *support* of 80% and a *confidence* of 95% were set as minimum thresholds. The thresholds mean that, for each generated association rule, at least 80% of all the data records under analysis contain both premise and conclusion; and the probability that a premise's emergence leads to a conclusion's occurrence is 95% or more. In addition, the minimum threshold of *lift* value was set 1 to find positive correlations. The mining in the dataset_1 generated 476 rules (i.e., the rule set 1) and 43 parameters were involved.

Then, the association rules were mined in the dataset_2 and only the data records of these 43 parameters were used. Such mining generated 169 rules (i.e., the rule set 2). Among the generated rules, many of them are obvious and uninteresting; and truly interesting rules need to be further identified based on the knowledge of building engineering. Also, the two rule sets (i.e., the rule sets 1 and 2) were compared with each other. As a result, three potentially useful association rules were found and given in Table 6-1.

Table 6-1 Three best rules generated

No.	Premise	Conclusion	*Sup*	*Conf*	*Lift*	Dataset
Rule 1	TA_{IVah} [HIGH], TA_{IVac} [LOW]	F_{IV} [HIGH], TA_{Vac} [LOW]	0.81	0.99	1.21	1
Rule 2	F_{IV} [HIGH], TA_{Vac} [LOW]	TA_{IVah} [HIGH], TA_{IVac} [LOW]	0.81	0.99	1.21	1
Rule 3	$TA\mathrm{I}_{\mathrm{Vac}}$ [LOW]	TA_{IVah} [HIGH]	0.78	1.00	1.12	2

Clearly, the premise and conclusion of the first two rules are reversed, and thus shows that the following four facts frequently occurred at the same time in winter 2007:

(1) The fresh air temperature after the heating coil in the FHU 4 was ‘HIGH’,

(2) The fresh air temperature after the cooling coil in the FHU 4 was ‘LOW’,

(3) The fresh air fan frequency of the FHU 4 in the VA side was ‘HIGH’,

(4) The fresh air temperature after the cooling coil in the FHU 5 was ‘LOW’.

Also, *Rule 3* shows that the following two facts frequently occurred at the same time in winter 2009:

(1) The fresh air temperature after the cooling coil in the FHU 4 was ‘LOW’,

(2) The fresh air temperature after the heating coil in the FHU 4 was ‘HIGH’.

Based on the facts 1, 2, 5, and 6, it was observed that, in winter, the fresh air temperature in the FHU 4 usually increased first and then significantly decreased, which indicates a possible waste of energy. In order to illustrate this observation clearly, the screenshot of the FHU 4 control panel is shown in Figure 6-3. In this diagram, the components in Δ, □, ○, ∇ are the heat recovery (recuperation), heating coil, humidifier and cooling coil, respectively.

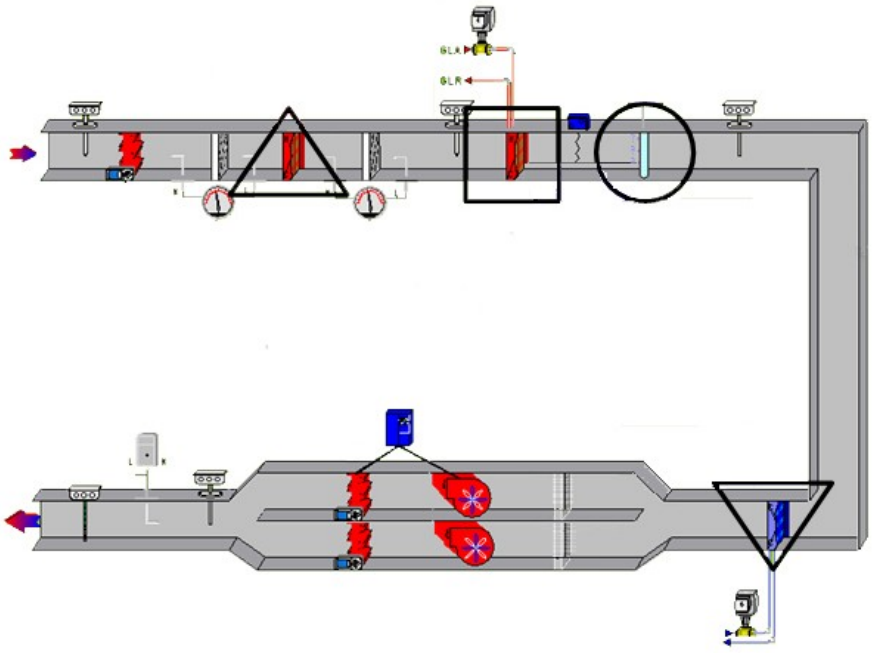

Figure 6-3 Screenshot of the FHU 4 control panel

The heating coil was always on while the cooling coil was always shut down in winter[2]. Hence, after the heating coil the temperature of fresh air drops only because of the humidifier that uses municipal water[2] at about 2°C. Site visit confirmed that this water was drained directly to sewage after humidification process. The heating and humidifying process is plotted in Figure 6-4 (left).

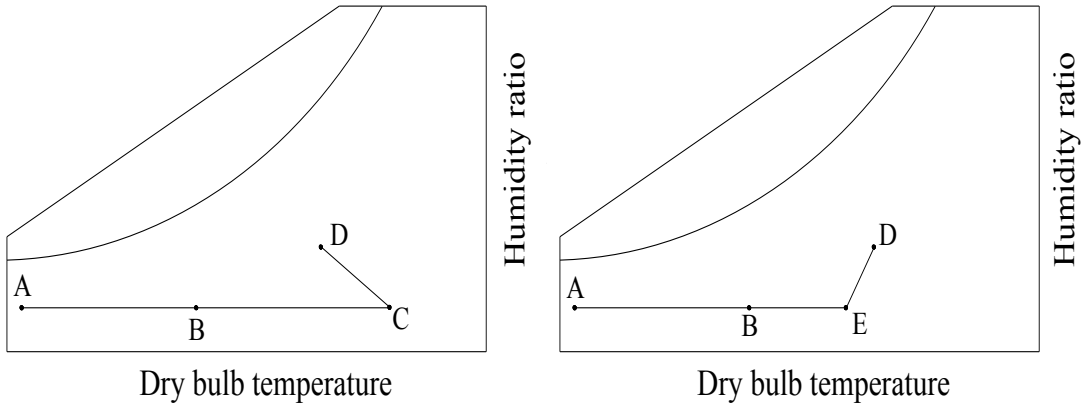

Figure 6-4 Heating and humidification processes in psychrometric chart

As seen in the left diagram of Fig. 6.4, outdoor air is at state point A. Process A-B represents sensible pre-heating and heat recovery, which can be characterized by a horizontal line. After this, heating and humidification are carried out successively, shown

[2] Information provided by the building operators.

[2] Information provided by the building operators.

as processes B-C and C-D. Based on the monitored data, the actual air temperature after the heating coil (point C) and the air temperature after the humidifier (point D) are plotted in Figure 6-5.

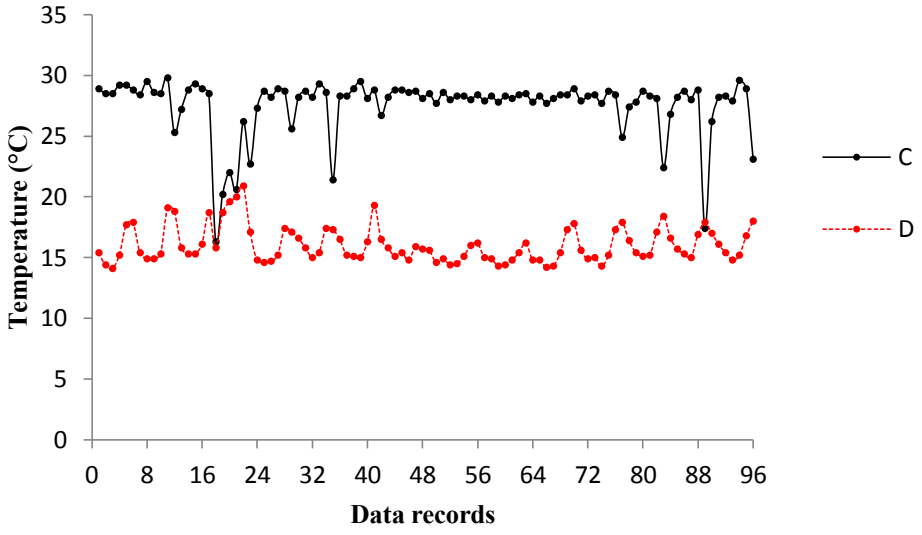

Figure 6-5 Air temperature after heating coil (state C) and humidifier (state D)

Figure 6-5 indicates that the air temperature after the heating coil is around 14°C higher than that after the humidifier. Clearly it is the low temperature of municipal water that caused the dramatic temperature drop (from state C to state D) in the conditioned fresh air, and such temperature drop can lead to a significant energy waste. That means the heat added to the fresh air during A to B process and B to C process is simply drained to municipal sewage after the humidifier.

One possible remedy for such an issue would be decreasing the air temperature after

the heating coil. More specifically, shift point C to the left (to point E), as shown in the right diagram of Figure 6-4. Correspondingly, one possible method in reality could be recycling and reusing (instead of discharging) the municipal water after it is warmed up after passing through the humidifier. In order to describe this process clearly, based on the monitored data and heat transfer theory, two schemas of hypothetical air/water temperature in the FHU 4 in winter before and after the remedy are given in Figures 6-6 and 6-7.

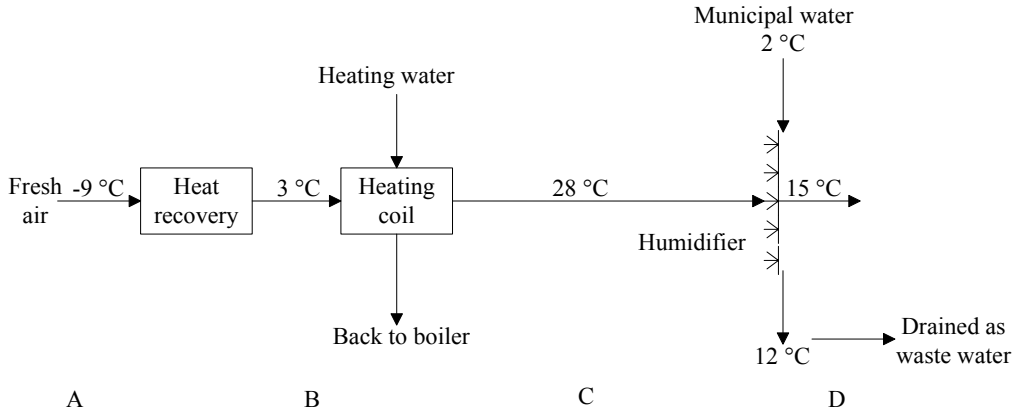

Figure 6-6 Hypothetical air/water temperature in the FHU 4 before the remedy

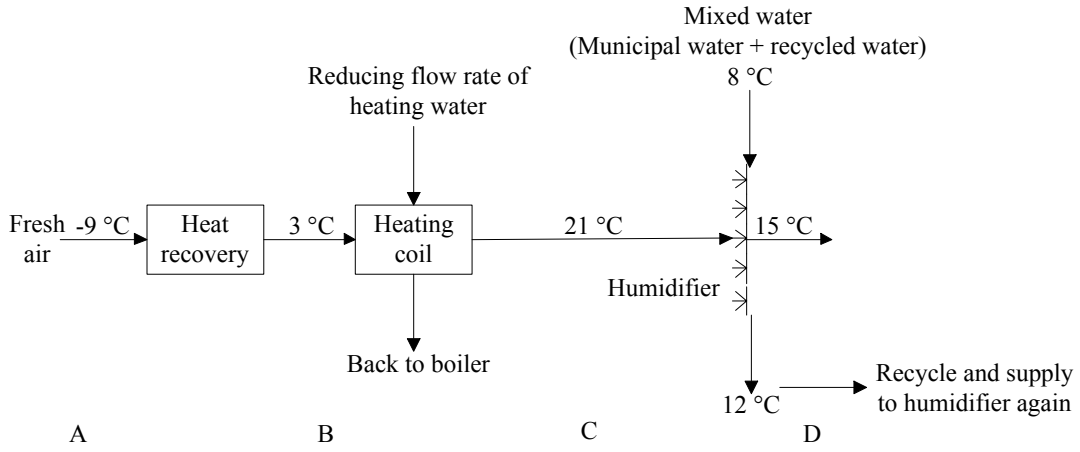

Figure 6-7 Hypothetical air/water temperature in the FHU 4 after the remedy

In Figure 6-6, the outdoor air temperature, air temperature after the heat recovery, air

temperature after the heating coil, and air temperature after the humidifier are assumed to be -9 °C, 3 °C, 28 °C and 15 °C, respectively. At the same time, municipal water before and after the humidifier are assumed to be 2 °C and 12 °C.

In Figure 6-7, the recycled high temperature municipal water (at 15 °C) and fresh municipal water (at 2 °C) could be mixed and then supplied to the humidifier again, considering the water loss during humidifying. The temperature of the mixed water is assumed to be at 8 °C and the water left the humidifier at 12 °C (or even higher). With this method, it would be enough to heat the fresh air up to a lower temperature (e.g., 21 °C as shown in Figure 6-7) instead of 28 °C in the heating coil. Accordingly, a huge amount of energy can be saved in the heating coil. However, it should be mentioned that it would be necessary to treat the water before it is reused[3] to prevent microbial issues.

6.4.2 ARM in winter in the dataset_1 and dataset_2

Association rule mining was also carried out for the dataset_1 and dataset_2. After experimenting with various combinations of *support* and *confidence* values, a *support* of 50% and a *confidence* of 80% were set as minimum thresholds. In addition, the minimum threshold of *lift* value was set 1 to find positive correlations. Specifically, association

[3] Through discussion with the building operators, this energy waste was confirmed and they planned to fix this problem using an appropriate method.

rules were first mined in the dataset_1. Such mining generated 461 rules (i.e., the rule set 3), and 32 parameters were involved in these rules. Then, association rules were mined in the dataset_2 and only the data records of these 32 parameters were used. Such mining generated 262 rules (i.e., the rule set 4). After that, the two sets of generated rules were compared with each other to further identify truly interesting rules. As a result, the obtained interesting rules were grouped into three categories in order to discover useful knowledge, as follows:

Category 1: same rules generated in the both datasets

Table 6-2 Four rules in Category 1

No.	Premise	Conclusion	*Sup*	*Conf*	*Lift*	Dataset
Rule 1	Q_{I1} [LOW]	Q_{II1} [LOW]	0.52	0.98	1.70	1
Rule 2	Q_{I1} [LOW]	Q_{II1} [LOW]	0.55	1.00	1.63	2
Rule 3	Q_{I2} [LOW]	Q_{II2} [LOW]	0.52	0.97	1.70	1
Rule 4	Q_{I2} [LOW]	Q_{II2} [LOW]	0.57	0.95	1.65	2

From *Rules 1* and *2*, it can be observed that, the airflow rates of fan 1 in the FHU 1 and FHU 2 have a strong association and correlation. At the same time, *Rules 3* and *4* show that the airflow rates of fan 2 in the FHU 1 and FHU 2 also have a strong association and correlation (this is reasonable since the two fans in the same FHU are identical and controlled by one variable speed drive (VSD)). Therefore, it can be inferred

that the total airflow rates of the FHU 1 and FHU 2 are strongly associated and correlated.

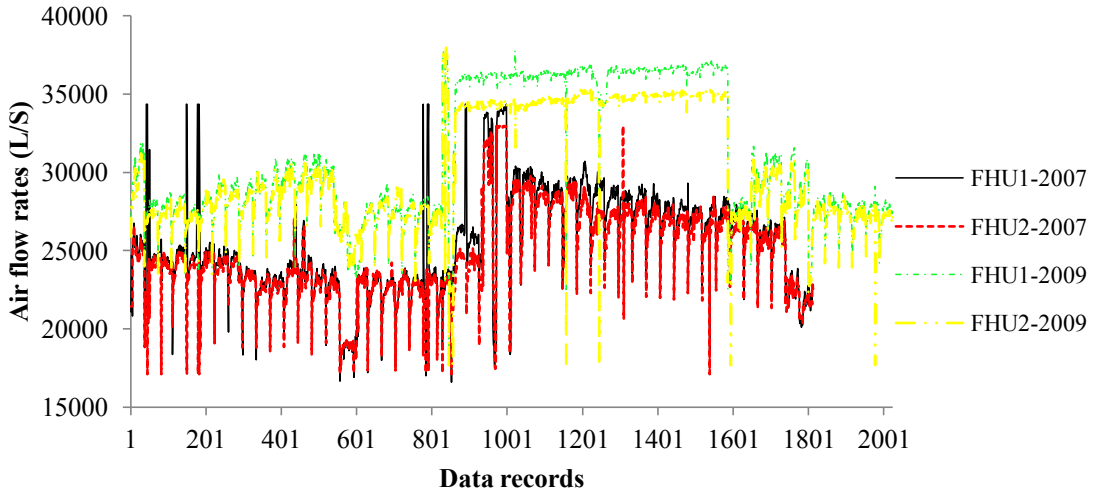

Figure 6-8 Air flow rates of the FHUs 1 and 2 in the dataset_1 and dataset_2

The airflow rates of the FHUs 1 and 2 in both dataset_1 and dataset_2 are plotted in Figure 6-8. It can be seen that the variation of airflow rates of these two FHUs follows the same trend. Furthermore, the values of airflow rates between these two FHUs are close to each other in both datasets. This indicates that the total airflow rates of the FHU 1 and FHU 2 are always strongly associated and correlated. Accordingly, if a continuous significant difference between them is observed, it can be inferred that either of the FHUs could have a fault. Therefore, the rules can help to understand FHU operation and also be applied to online fault detection.

Category 2: similar rules generated in both the datasets but are opposite in premise/conclusion

Table 6-3 Six rules in Category 2

No.	Premise	Conclusion	*Sup*	*Conf*	*Lift*	Dataset
Rule 1	Q_{V1} [LOW]	Q_{IV1} [LOW]	0.59	0.92	1.49	1
Rule 2	Q_{V1} [HIGH]	Q_{IV1} [LOW]	0.51	0.81	1.31	2
Rule 3	Q_{V2} [LOW]	Q_{IV2} [LOW]	0.57	0.91	1.50	1
Rule 4	Q_{V2} [HIGH]	Q_{IV2} [LOW]	0.54	0.99	1.31	2
Rule 5	Q_{IX3} [LOW]	TA_{IXbri} [HIGH]	0.60	0.82	1.12	1
Rule 6	Q_{IX3} [HIGH]	TA_{IXbri} [HIGH]	0.52	0.90	1.51	2

Six potentially useful rules in Category 2 are found and given in Table 6-3. *Rules 1* and *2* show that, between these two years, the airflow rates of fan 1 in the FHU 4 and FHU 5 have opposite associations and correlations. Similarly, *Rules 3* and *4* can also be explained.

In order to provide an insight into the association opposition, the airflow rates of fan 1 in the FHUs 4 and 5 in these two years are plotted in Figures 6-9 and 6-10, respectively. Considering that fan 1 and fan 2 in the same FHU are identical and controlled by the same VSD, their airflow rates are approximately the same, and thus only the airflow rate of the fan 1 is plotted.

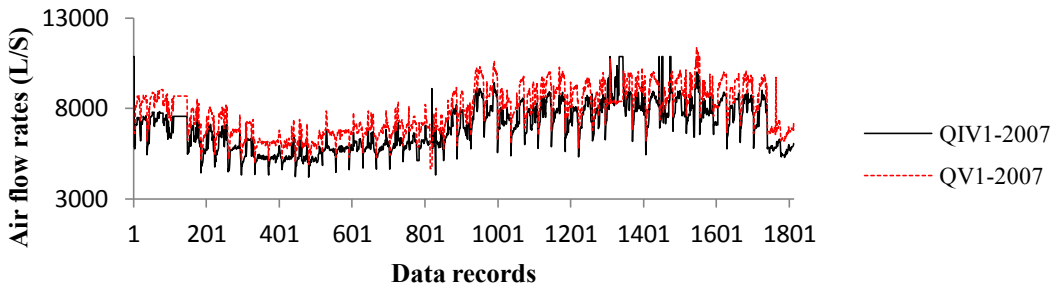

Figure 6-9 Air flow rates of fan 1 in the FHUs 4 and 5 in dataset_1

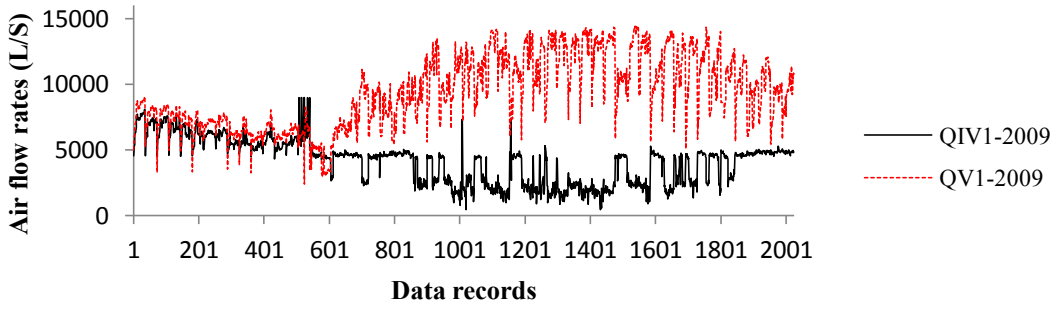

Figure 6-10 Air flow rates of fan 1 in the FHUs 4 and 5 in dataset_2

Figure 6-9 clearly shows that the values of air flow rates of fan 1 in these two FHUs are very close in 2007. This is reasonable since these two FHUs are identical, and clearly their airflow rates should always be almost the same. However, Figure 6-10 shows that, in 2009, the airflow rates of fan1 in the FHU 5 are much larger than that in the FHU 4 most of the time. Accordingly, it can be inferred that a fan fault occurred in the FHU 4 in 2009. Therefore, the rules can be used as a guide of fault diagnosis on the fans and FHUs.

Based on *Rules 5* and *6*, it can be found that these two rules' premises (i.e., the

airflow rate of fan 3 in the EHU 2) are opposite.

Figure 6-11 shows the screenshot of the EHU 2 control panel. Clearly, exhaust air from different parts of the VA part will be mixed in duct 4 before being distributed to the three exhaust air ducts (refer to 1, 2 and 3 in this diagram) and the three fans (refer to three yellow circles in this diagram). A further analysis of operational data on these three fans in both years shows that two of them were always turned on to extract exhaust air while the other one was turned off. Moreover, two different control strategies were implemented in the two different years respectively: in 2007, the fans 1, 2, and 3 were turned off alternatively; in 2009, the fan 2 was always turned off while the fans 1 and 3 were always turned on. However, from the point of view of energy consumption, there is no difference between these two strategies, and it is highly desirable that a new control strategy can be proposed to save energy. Given that these three fans are identical and controlled by individual VSD, one possible energy-saving method is to use all these three fans instead of two of them. A comparison between the current and proposed strategy is made to show the energy conservation. For current strategy, assume the actual air flow rate of each fan is M, the actual fan speed is V, and the actual power required by each fan is P. Table 6-4 shows the results of comparison between the two strategies.

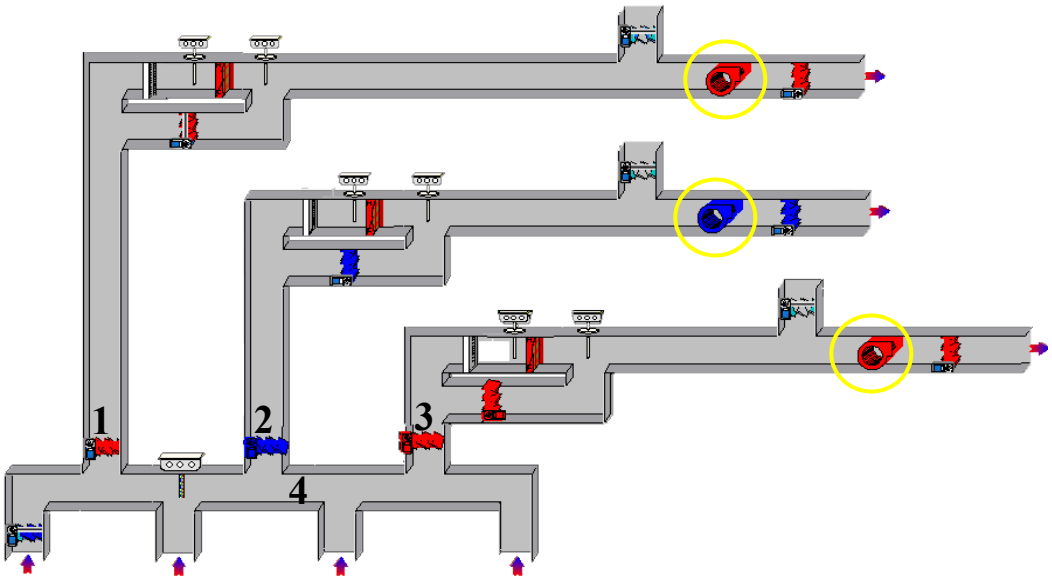

Figure 6-11 Screenshot of the EHU 2 control panel

Table 6-4 Comparison between the two control strategies

Strategy	Number of fans used	Air flow rate of each fan	Total air flow rate	Fan speed	Power required by each fan	Total power required
Current	2	M	$2M$	V	P	$2P$
Proposed	3	$2M/3$	$2M$	$2V/3$[a]	$8P/27$[b]	$8P/9$

[a] According to the *fan laws*, the capacity is directly proportional to the fan speed.

[b] According to the *fan laws*, the power required is proportional to the cube of fan speed.

From Table 6-4, it is obvious that $(2P\text{-}8P/9) = 10P/9$ can be saved if the proposed strategy is used. However, before this strategy is adopted in practice, it should be checked whether the fans will operate in the range of high efficiency, but not the dangerous unstable (surge) region at low air flow rates.

Category 3: rules generated in only one dataset (either dataset_1 or dataset_2)

Table 6-5 One rule in Category 3

No.	Premise	Conclusion	*Sup*	*Conf*	*Lift*	Dataset
1	F_{VI} [HIGH]	F_{VII} [HIGH]	0.60	0.97	1.60	1

One potentially useful rule in Category 3 was found and given in Table 6-5. The rule shows that the fan frequency in the RHU 1 and RHU 2 has a strong association and correlation. The frequency of the two fans is plotted in Figure 6-12, and it can be seen that F_{VI} is almost equal to F_{VII} all the time. Given that the RHU 1 and RHU 2 are identical, it can be inferred that these two RHUs' air flow rates (i.e., Q_{VI} and Q_{VII}) should be approximately identical. Accordingly, there should exist a strong association and correlation between Q_{VI} and Q_{VII}. However, no rule between Q_{VI} and Q_{VII} has been found in both dataset_1 and dataset_2. For this reason, air flow rates of the fan in the RHUs 1 and 2 in the dataset_1 are plotted in Figure 6-13. Clearly, a significant difference can be found between Q_{VI} and Q_{VII}, which indicates that either RHU 1 or RHU 2 has a fault. Further, data shows that the RHU 1 did not operate in 2009 (Q_{VI} is zero in the dataset_2). Therefore, it can be concluded that the RHU 1 has a fault.

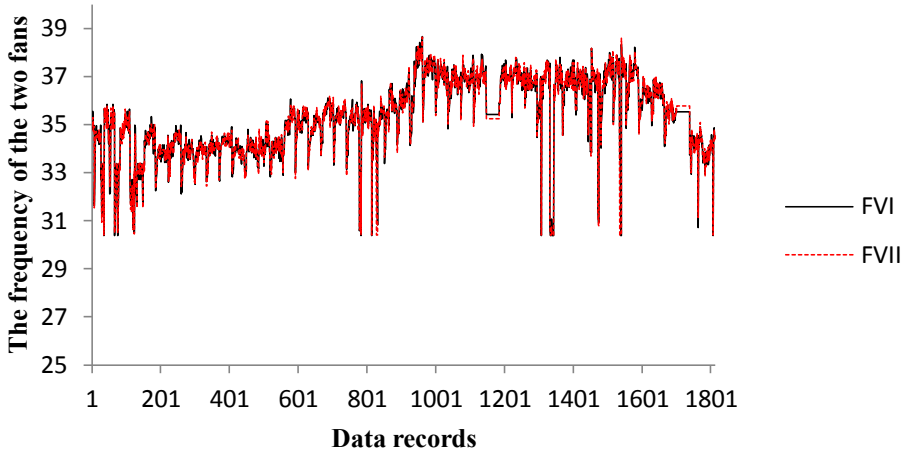

Figure 6-12 Frequency of VSD on the fan in the RHU1 and RHU2 in dataset_1

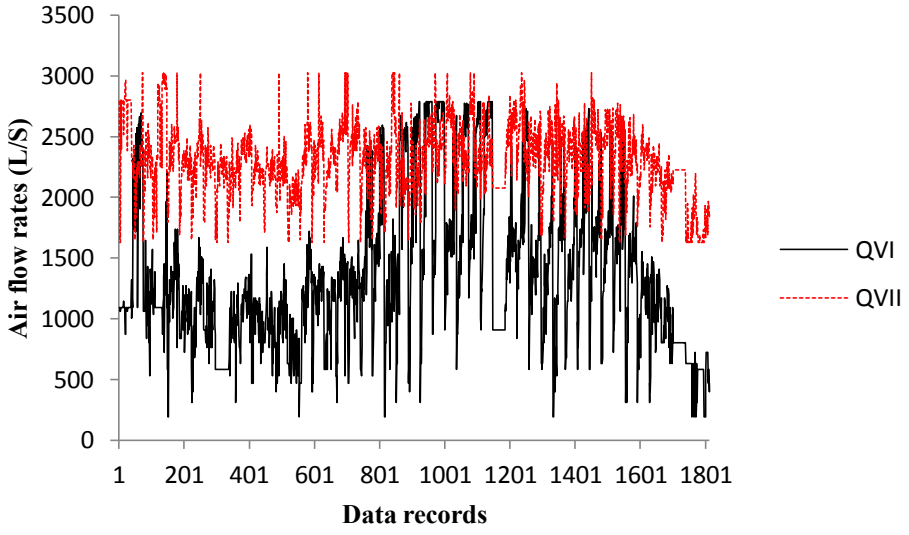

Figure 6-13 Air flow rates of the fan in the RHUs 1 and 2 in dataset_1

6.4.3 Association map

Besides association rules in the form of text, RapidMiner also provides a graphical view of an association map, representing all generated association rules. For simplicity, the association map in the dataset_2 instead of the dataset_1 is given in Figure 6-14,

considering that only the parameters showing up in both the rule set 3 and the rule set 4 are involved.

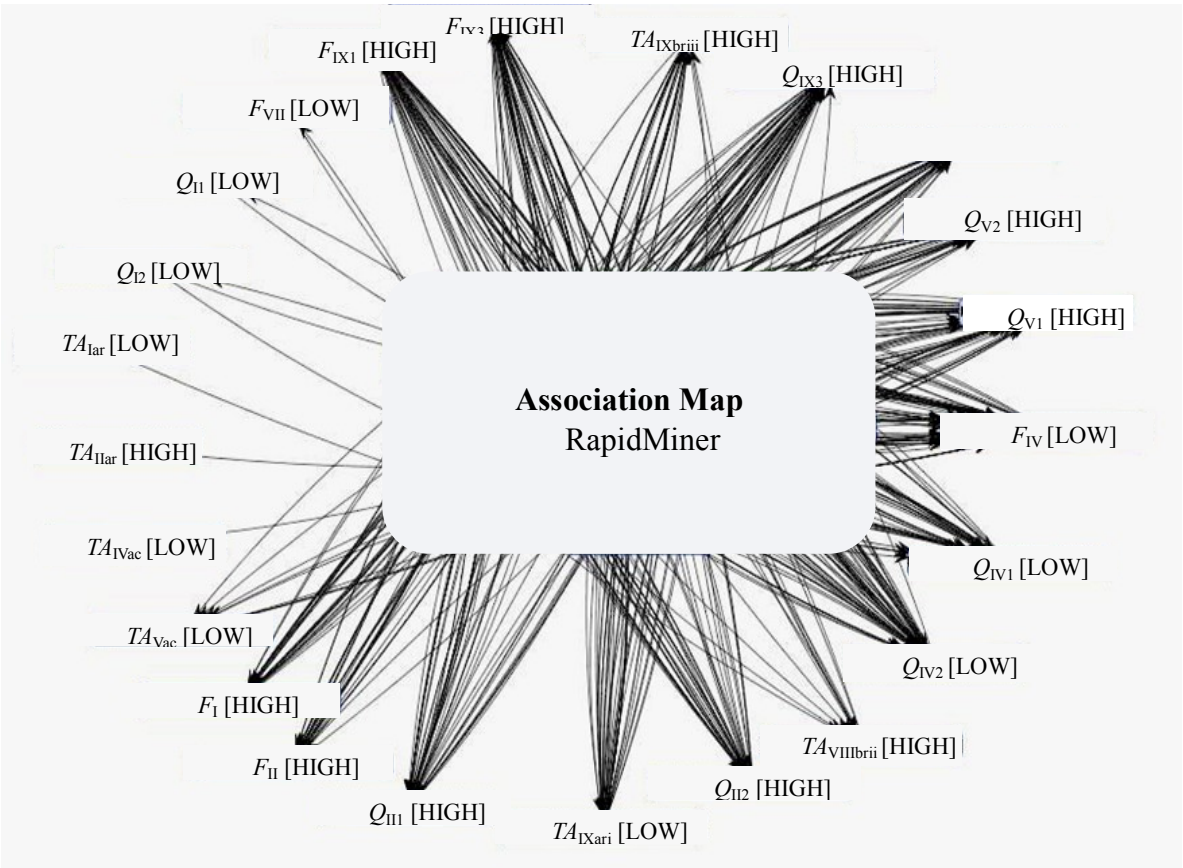

Figure 6-14 Association map in the dataset_2 provided by RapidMiner

In this map, each line represents one association rule, and thus the amount of lines quantitatively indicates the amount of associations between various parameters. Moreover, an arrow towards the parameter shows that this parameter appears in the conclusion of the association rule, and vice versa.

The map provides a holistic pattern of associations between various parameters.

Clearly it can be seen that there exists a significant difference between the parameters on the amount of associations with other parameters. For example, TA_{Iar} and TA_{IIar} (i.e., fresh air temperature after the recuperation glycol in the FHUs 1 and 2) have only one association with other parameters and both of them appear in the premise. This indicates that these two parameters' values may be purely random or remain relatively stable throughout the whole winter and thus no association with other parameters can be found. It may have occurred since these two parameters are partly decided by outdoor air temperature, which is uncontrollable and relatively irregular. On the contrary, Q_{IV2} (i.e., the fresh air flow rate of fan 2 in the FHU 4) has the most associations with other parameters, and appears in both premises and conclusions. This indicates the parameter has the highest possibility of influencing or being influenced by other parameters and thus deserves extra attention.

In addition, between similar parameters (e.g., air flow rates of two fans in the same FHU), difference in the amount of associations with other parameters should not be huge. However, it is noticed that, between TA_{IVac} and TA_{Vac} (i.e., the fresh air temperature after the cooling coil in the FHUs 4 and 5), such difference is significant: TA_{IVac} only has one association with other parameters while TA_{Vac} has eight. This implies that the FHU 4 may have a fault. Accordingly, data analysis was performed on various parameters of the FHU 4; and the air flow rates of fans 1 and 2 in the FHU 4 are plotted in Figure 6-15. Clearly, the air flow rates between these two fans are completely different most of the time. Considering

fan 1 and fan 2 in the same FHU are identical and controlled by the same VSD, it can be inferred that, either the fan 1 or the fan 2 (or both of them) in the FHU 4 has a fault. This conclusion is in accordance with the conclusion drawn from *Rules 1* to *4* in Category 2 (Section 6.4.2).

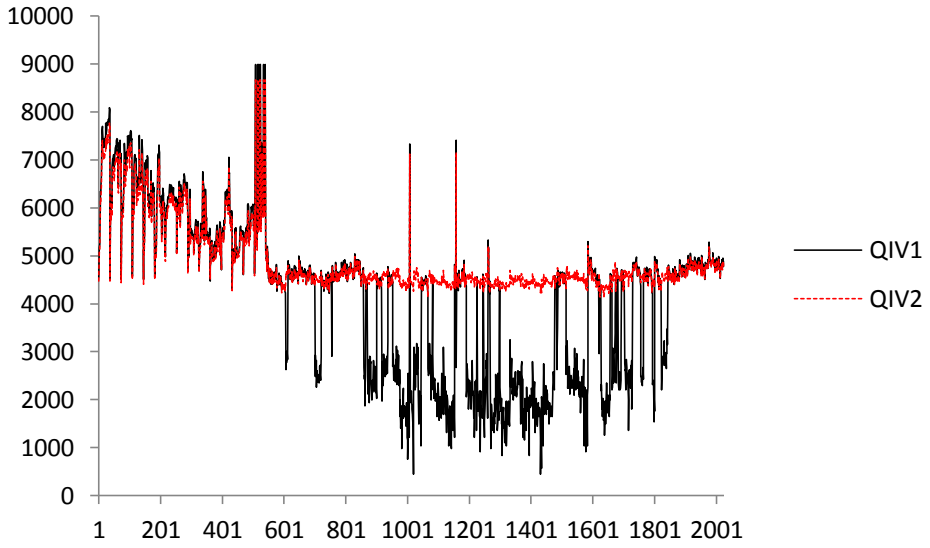

Figure 6-15 Air flow rates of fans 1 and 2 in the FHU4 in the dataset_2

The acquired knowledge could help building operators and owners better understand HVAC system operation and detect faults.

6.5 Summary and conclusions

In this chapter, a methodology is proposed for examining all the associations and correlations between building operational data. Accordingly, useful knowledge will be uncovered to help improve HVAC system performance and reduce energy consumption.

The methodology is based on a basic data mining technique: association rule mining. In order to use this methodology, two-year building operational data needs to be collected. Data pre-processing should be performed before the ARM to remove outliers, so as to improve the quality of data and, consequently, the mining results. Furthermore, to take complete advantage of building operational data, data in different period length (e.g., both a day and a year) should be mined. Also, the obtained associations and correlations in different years should be compared between each other.

In order to demonstrate its applicability, this methodology was applied to the EV building located in Montreal, which is very cold in winter. Accordingly, the winter data of the air-conditioning system in this building in both 2007 and 2009 was mined. A waste of energy in the air-conditioning system was identified through mining association rules for the coldest day. Also, based on the comparison between winter association rules in the different years, possible faults in equipment were detected, and a low/no cost strategy for saving energy in system operation was proposed. Moreover, the association map was used to provide a holistic view of all the generated rules. This map could help explain how various parameters associate one with each other, and detect faults in equipment.

The proposed methodology allows for addressing the special challenges caused by the complexity of large volume of building operational data. By using this methodology, building operators and owners can discover all the useful associations and correlations between building operational data. Based on domain expertise, they can translate the

obtained associations and correlations into useful knowledge, thereby better understanding building operation, identifying energy waste, detecting faults in equipment, and proposing low/no cost strategies for saving energy.

7. A METHODOLOGY FOR IDENTIFYING AND IMPROVING OCCUPANT BEHAVIOR IN RESIDENTIAL BUILDINGS

7.1 Introduction

Among various factors influencing residential building energy consumption, occupant behavior plays an essential role and is difficult to investigate analytically due to its complicated characteristics. Note that here occupant behavior refers to activities that have a direct or indirect impact upon building energy consumption. For example, occupants turn on/off lights, TV sets, computers, microwave ovens, and so on. Commonly such behavior is associated with various household appliances and thus can be deduced indirectly from corresponding end-use loads. For example, the total daily (or monthly, annual) lighting energy consumption in a residential building qualitatively indicates the duration of lighting usage in this day (or month, year). Accordingly, any improvement in the occupant behavior leads to the reduction of the residential building energy consumption. Therefore, it is necessary to develop a methodology to help occupants identify and improve their behaviour that needs to be modified.

This chapter reports the development of a rational methodology for identifying and improving occupant behavior in residential buildings, based on an analysis of collected

data and information. In particular, feasible recommendations are made for assisting occupants to modify their behaviour so as to reduce energy consumption.

7.2 Methodology

A methodology is proposed for efficiently improving occupant behavior in residential buildings, and evaluating the energy-saving potential resulting from these modifications. As mentioned previously, end-use loads are used to deduce user activities indirectly. Specifically, these loads are used to map onto occupant behavior at two levels, as shown in Figure 7-1.

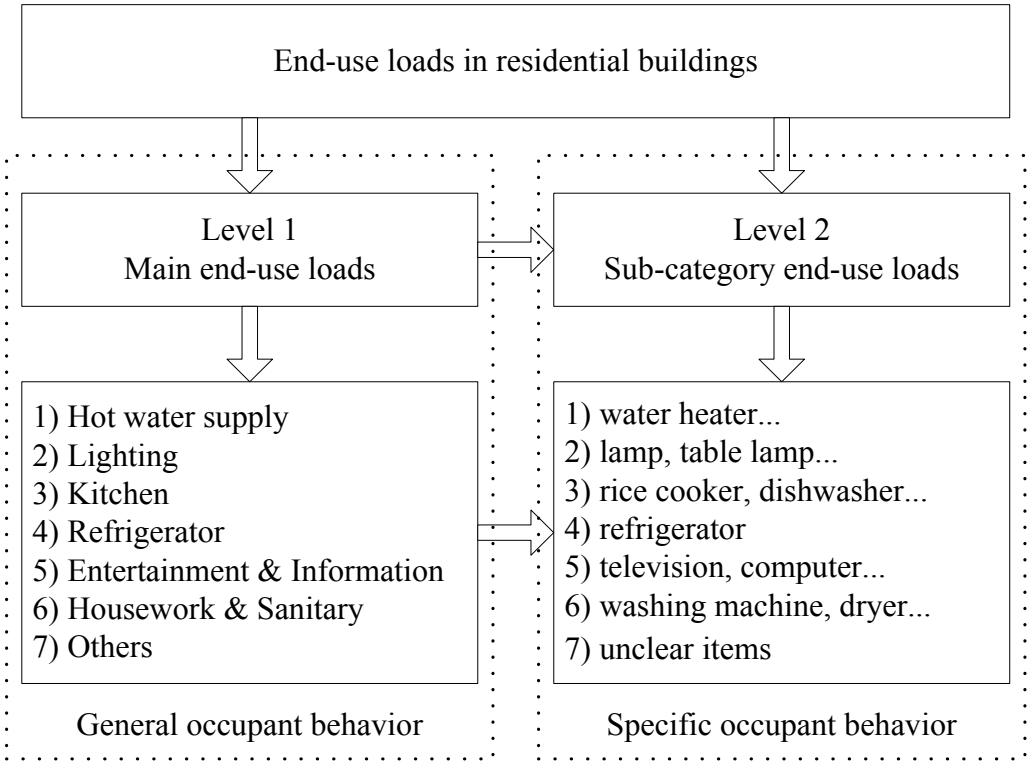

Figure 7-1 Two-level end-use loads

Level 1 loads are divided into seven main end-use loads, each of which can be further divided into various end-users in level 2. The seven end-use loads in level 1 are assumed to be non-weather-dependent (Zmeureanu et al., 1999), due to the fact that the usage of these appliances (i.e., lighting, refrigerators, etc.) is mainly determined by occupants' presence and behaviour. It should be mentioned that, the level 2 end-users are not fixed in different residential buildings since commonly different families have different household appliances. The level 1 and level 2 loads are mapped onto general occupant behavior, such as activities associating with lighting and hot water supply, and specific occupant behavior, such as the use of computers and washing machines.

For demonstration purposes, a group of buildings is used to show the practical application of this methodology. Recommendations for improving occupant behavior are provided for a selected building (*case building*) within this group. The methodology is briefly described as follows:

(1) Identify energy-inefficient general occupant behavior in the *case building*,

(2) Identify a *reference building* for the *case building* to evaluate its energy-saving potential, and further determine its energy-inefficient general occupant behavior by comparison with the *reference building*, and

(3) Identify energy-inefficient specific occupant behavior in the *case building*.

The proposed methodology can be demonstrated in a five-step process, as shown in Figure 7-2.

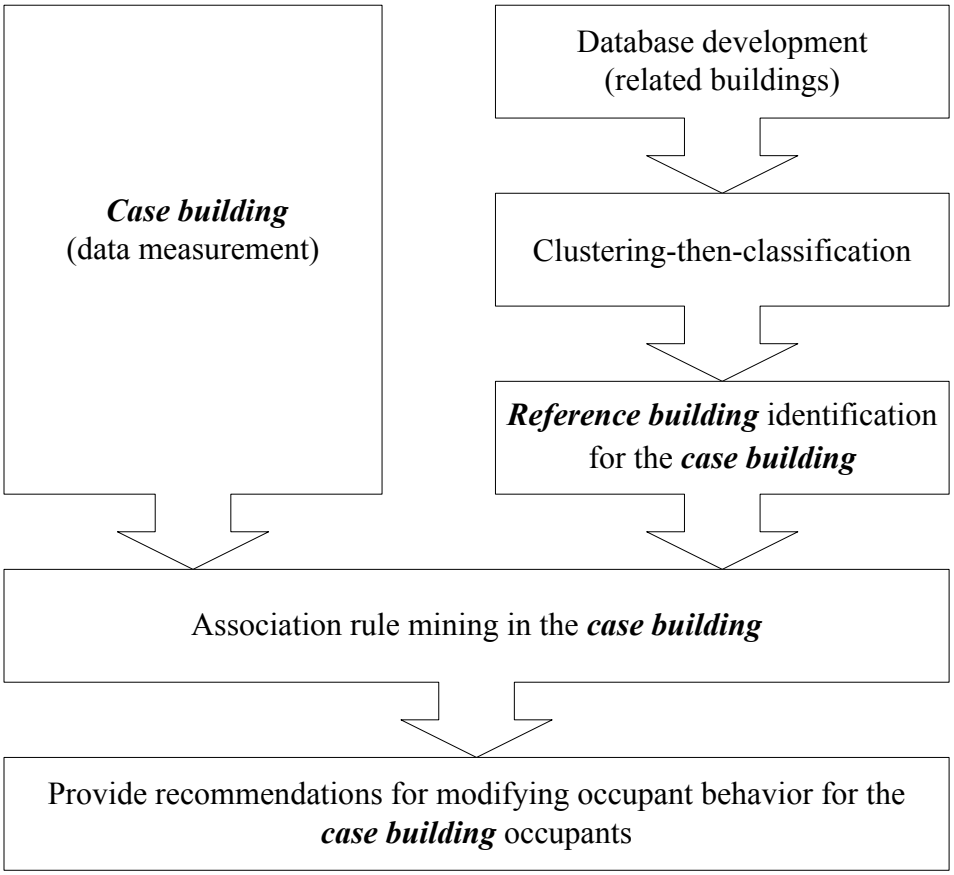

Figure 7-2 Methodology of evaluating and efficiently improving occupant behavior in the case building

Each step in this methodology is briefly explained as follows:

(1) First, a database should be developed based on the collection of measured data for the *case building* and other related buildings (e.g., buildings selected in the same city or country). The daily (or hourly) level 2 end-use loads should be measured, and the level 1 end-use loads can be accumulated based on the level 2 data. The database should also contain information about building-related parameters, such

as floor area and number of occupants.

(2) Through clustering analysis, all the related buildings in the database are clustered into different groups in terms of the level 1 loads (for each main end-use load, the annual per capita end-use loads is used for comparison). Accordingly, general occupant behavior in different buildings in the same group has a high similarity, but is quite different from that in other groups. Specifically, comparing with occupants in other clusters, on average each occupant in the same cluster consumes similar amounts of energy each year in terms of the seven level 1 end-use loads. Note that these seven loads are taken into consideration separately but simultaneously. Consequently, by comparing with other clusters, the characteristics of occupant behavior in each cluster can be identified. Such information can help building occupants to evaluate their own behavior among all the building owners in the database, thereby identifying general occupant behaviour which results in inefficient use of energy. Then, data classification based on the generated clusters is performed, and specifically, a decision tree (Han et al., 2006) is developed. By using the generated decision tree, a building can be assigned to a specific cluster, provided its level 1 loads are available. In particular, once the *case building* has been assigned to a cluster, its general energy-inefficient occupant behaviour can be determined. It should be mentioned that, the decision tree was selected and used in this study due to the fact it can

provide useful information which can help to understand the role of building occupant behavior in improving energy saving.

(3) Among the related buildings in the database, a *reference building* (RB) is identified for the *case building* to evaluate its energy-saving potential due to the occupant behavior modification. The RB is selected from the same cluster as the *case building* so that both of them have similar holistic occupant behavior patterns. The comparison with the RB also shows the *case building* occupants which general occupant behavior still needs to be modified.

(4) After identifying the energy-inefficient general occupant behavior through clustering analysis and RB identification, it is necessary for the *case building* owner to know which specific activities and corresponding appliances deserve extra attention. Therefore, association rules are mined to identify the associations and correlations between various user activities in the *case building*, in order to highlight energy-saving opportunities.

(5) Recommendations for energy-efficient activities are provided for the *case building* occupants, so that they can modify their behavior.

7.3. Reference Building (RB) identification

The steps in identifying a RB for the *case building* are explained as follows.

RB is normally utilized as a benchmark for comparison and the method of defining a RB depends on the purpose of study. In this study, the RB was defined to evaluate the

energy-saving potential due to occupant behavior modification in the *case building*, and identify occupant behavior that needs to be improved. Therefore, the definition of RB for the *case building* should comply with the following two rules:

Rule 1: The holistic occupant behavior patterns in RB and the *case building* should be as similar as possible. Different residential building occupants normally have different lifestyles and behavior patterns. In general, it is difficult for building occupants to make dramatic lifestyle changes in order to reduce energy consumption. Hence, among the related buildings in the database, buildings with similar occupant behavior patterns should be considered when evaluating the energy-saving potential for the *case building*. This implies that potential RB candidates should be chosen from buildings in the same cluster as the *case building*, since occupant behavior in the same cluster has a high similarity in comparison to one another, but is quite dissimilar to that in the other clusters.

Rule 2: Among all the potential RB candidates, the selected RB should have the highest similarity to the *case building* in terms of building-related parameters, such as outdoor temperature and floor area. This can also improve the reliability of comparative results between the two buildings. Euclidean distance can be used to define the similarity. With consideration of the two rules, RB identification for the *case building* consists of the following steps:

Step 1 Assign the '*case building'* to a cluster according to the level 1 loads and the generated decision tree;

Step 2 Calculate the total energy consumption (i.e., the sum of the seven main end-use loads) in the *case building* and other buildings in the same cluster. Rank the total energy consumption in all these buildings; and

Step 3 Identify the RB. Buildings in the same cluster with lower total energy consumption than the *case building* are used as potential RB candidates. Then, based on building-related parameters and Euclidean distance, the most similar building to the *case building* among the candidates can be found. This building is identified as RB for the *case building*.

7.4 Data pre-processing

7.4.1 Case building selection

As mentioned earlier, for demonstration purposes, one building with the most comprehensive household appliances should be selected as the *case building*, and the remaining 66 buildings are used for both clustering-then-classification and RB identification. Data inspection indicates that a building located in Hokkaido has the most appliances, as shown in Table 7-1. This Table also shows some measured environmental parameters of this building such as indoor air temperature and humidity. These parameters will also be used in the ARM to analyze the associations between them and occupant behaviour.

Table 7-1 Appliances in the case building and environmental parameters used in ARM

No.	Appliances/ indoor parameters	No.	Appliances/ indoor parameters	No.	Appliances/ indoor parameters
1	Heating boiler	16	TV (other rooms)	31	Living room temperature
2	Hot water boiler	17	TV (standby power)	32	Living room humidity
3	Kerosene heater	18	Video	33	Bedroom (1F) temperature
4	Ventilator	19	Phone	34	Master bedroom (2F) temperature
5	Air cleaner	20	Telephone handset	35	Total energy consumption
6	Lamp (1F[a*])	21	Iron	36	SHW
7	Lamp (2F[b*])	22	Vacuum cleaner	37	LIGHT
8	Table lamp	23	Washing machine (1F)	38	KITCH
9	IH heater	24	Washing machine (2F)	39	REFRI
10	Dishwashers	25	Living room outlet	40	E&I
11	Microwave, toaster, coffee	26	Rest room outlet (1F)	41	H&S
12	Bidet	27	Rest room outlet (2F)	42	OTHER
13	Boom box	28	Outdoor air temperature		
14	TV (Dining room)	29	Outdoor relative humidity		
15	TV (master bedroom 2F)	30	Outdoor air velocity		

[a*] first floor, [b*] second floor.

Table 7-2 shows the statistical data of the level 1 loads for the remaining 66 buildings. Clearly, it can be seen that each main end-use load is spread over a wide range, which implies a fairly large energy-saving potential by improving occupant behavior.

Table 7-2 Statistical data of the seven main end-use loads for the 66 buildings

(unit: MJ per capita per year)

End-use load	Min	Max	Average	Standard deviation
SHW	994.945	11649.175	4695.497	2616.451
LIGHT	130.372	2938.521	1311.695	846.283
KITCH	110.761	5321.785	971.773	786.056
REFRI	390.136	2667.98	883.033	439.375
E&I	106.254	2301.679	727.136	480.946
H&S	64.137	2102.968	400.303	385.46
OTHER	55.259	2374.798	738.422	564.375

7.4.2 Data transformation for cluster analysis

Before performing the cluster analysis on the level 1 data, it should be noted that the loads, which were mapped onto various corresponding user activities, have different ranges. Moreover, the activities were considered to be of equal importance in this study. In order to prevent the loads with large ranges from outweighing those with comparatively smaller ranges, min-max normalization was applied before clustering the buildings in terms of the seven main end-use loads, which were introduced in Section 5.2.1.

In this study, the new range is defined as [0, 1]. Table 7-3 shows the statistical data of the level 1 loads for the remaining 66 buildings after min-max normalization.

Table 7-3 Statistical data after normalization

End-use load	Min	Max	Average	Standard deviation
SHW	0	1	0.347	0.246
LIGHT	0	1	0.421	0.301
KITCH	0	1	0.165	0.151
REFRI	0	1	0.216	0.193
E&I	0	1	0.283	0.219
H&S	0	1	0.165	0.189
OTHER	0	1	0.295	0.243

7.4.3 Removal of outliers for conducting ARM in the case building

As introduced in Section 6.3, removal of outliers plays a crucial role in preparing for the ARM, since outliers produce a large measure of skewness and have a significant influence on the partition of attribute values into different intervals. In this study, the method based on the lower quartile (Q_1) and the upper quartile (Q_3) of the standard boxplot was used (see Section 6.3). Moreover, in order to perform the ARM, the value of quantitative attributes generally needs to be classified into categorical values. Considering that most attributes used in the ARM in this study are end-use electricity loads, a two-interval scale (i.e., HIGH and LOW) was applied to represent high and low energy consumption using the same classification method in Section 6.3. Such high and low energy consumption can then be qualitatively mapped onto *energy-inefficient* and *energy-efficient* occupant behavior. It should be mentioned that HIGH and LOW quite possibly, but do not necessarily, correspond to *energy-inefficient* and *energy-efficient*

occupant behaviour in practice. For example, less energy efficient appliances will also cause higher energy consumption. However, given that energy-inefficient behaviour will waste energy and normally cause high energy consumption, such mapping was still used in this study. Consequently, the results need to be carefully analyzed and energy-inefficient behaviour should be eventually identified based on practical occupant behaviour patterns.

With consideration of the seasonality of occupant behavior, the ARM was performed based on seasonal data instead of annual data in this study for demonstration purposes. Given that the *case building* is located in Hokkaido, the coldest area in Japan, the winter data in 2003 was mined to generate association rules. Figure 7-3 shows the distribution of two intervals of all the ARM attributes after the removal of outliers. Note that the numbers in the abscissa represent the ARM attributes, and correspond to the number in Table 7-1. Clearly, it can be observed that most of the percentages range from 30% to 70%, indicating a roughly uniform distribution.

Commonly used ARM algorithms include the Apriori algorithm and the frequent-pattern growth (FP-growth) algorithm (Han et al., 2006). In this study, we employed the FP-growth algorithm to mine association rules due to its high efficiency and wide applicability. At the same time, the *K*-means algorithm and the decision tree method were used in the cluster analysis and data classification, respectively. In addition, the open-source data mining software RapidMiner was used as data mining tools to

analyze the data.

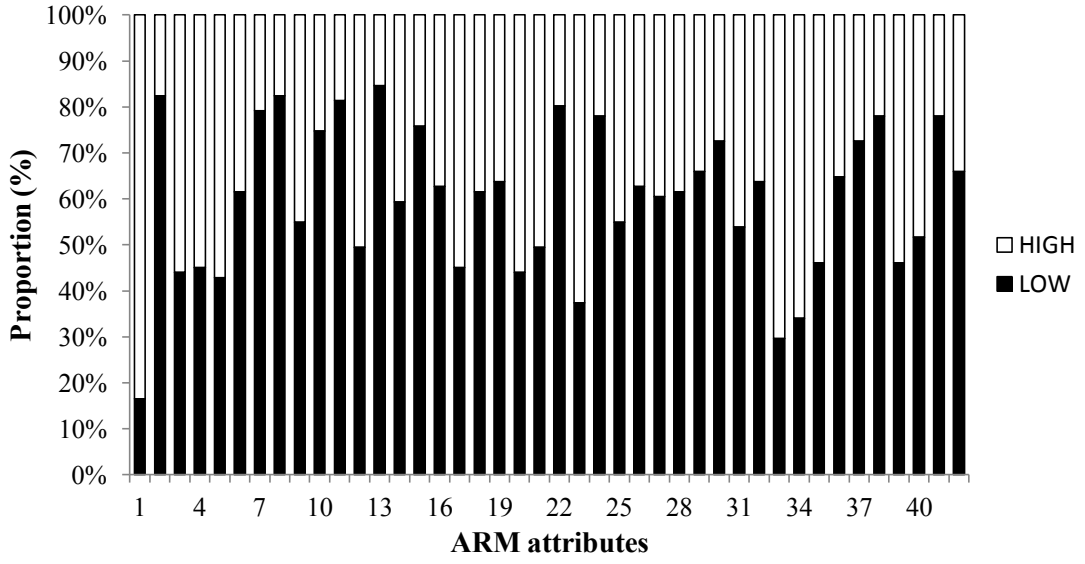

Figure 7-3 Distribution of two intervals of all ARM attributes after the removal of outliers

7.5 Results and Discussion

7.5.1 Clustering-then-classification

Clustering results

After data pre-processing, the cluster analysis was conducted for the 66 buildings using RapidMiner. With consideration of the size of the database, four clusters were determined by the *K*-means algorithm and the performance vector (Davies Bouldin index, DBI). The results of the cluster analysis are given in Table 7-4. Cluster centroids, which

represent the mean value for each dimension, were used to characterize building occupant behavior in the four clusters. For example, in comparison with building occupant behavior in the other clusters, user activities in cluster_2 caused medium energy consumption in supply hot water (the cluster centroid of SHW in this cluster is 0.440, which is of medium value among the four clusters), high energy consumption in lighting, medium energy consumption in kitchen, etc. Moreover, cluster_2 has significantly higher energy consumption for lighting; this indicates that, in general, building owners in cluster_2 should give primary consideration to the activities related to lighting in order to save energy. Similarly, other clusters can be explained. It should be noted that nearly half of the data records (44%) were grouped into cluster_1, which represents low energy consumption in most of the main end-use loads. A possible explanation for this is that a good portion of Japanese families have a high degree of awareness regarding energy-savings. In addition, among the seven attributes and four clusters, H&S has the largest maximum/minimum ratio (0.509/0.088 = 6.5), while KITCH has the lowest maximum/minimum ratio (0.268/0.144 = 1.91). This indicates that occupant behavior related to H&S differs significantly between the four clusters; and deserves extra attention in occupant behavior improvement; on the contrary, the total energy consumption caused by KITCH-related user activities has a narrow gap between different clusters, which implies relatively small energy-saving potential for modifying such kind of activities.

Table 7-4 Centroid of each cluster and statistics on the number and percentage of instances assigned to different clusters

Attribute	Cluster_1	Cluster_2	Cluster_3	Cluster_4
SHW	0.266	0.440	0.738	0.215
LIGHT	0.262	0.881	0.291	0.288
KITCH	0.144	0.181	0.268	0.140
REFRI	0.119	0.255	0.372	0.296
E&I	0.218	0.169	0.572	0.403
H&S	0.088	0.167	0.509	0.150
OTHER	0.136	0.430	0.231	0.500
Clustered buildings and proportion	29 (44%)	16 (24%)	7 (11%)	14 (21%)

Table 7-5 shows the number of buildings in various districts in each cluster. Clearly, the distribution of buildings in various districts is roughly even, especially in cluster_1 and cluster_4. Such a distribution indicates that the attributes in the cluster analysis are not dependent on weather (otherwise buildings in the same districts would tend to be grouped together), which is consistent with the assumption that the seven main end-use loads in clustering analysis are non-weather-dependent components.

Table 7-5 The number of buildings in various districts in each cluster

cluster	Hokkaido	Tohoku	Hokuriku	Kanto	Kansai	Kyusyu
cluster_1	6	3	7	3	5	5
cluster_2	0	4	0	8	2	2
cluster_3	1	2	4	0	0	0
cluster_4	3	2	1	1	5	2

Classification by decision tree

(1) Generation of decision tree

After the four clusters were generated, a decision tree was constructed to assign buildings to a specific cluster provided their main end-use loads are available, as shown in Figure 7-4. C4.5 algorithm was used in RapidMiner to build the decision tree.

The decision tree includes a total of 19 nodes among which 10 are leaf nodes. The colors in the leaf nodes indicate the purity of classification in the nodes. A pure color in a node implies that all the records in this node are correctly classified. Clearly, all the data records in the training dataset are correctly classified in this decision tree.

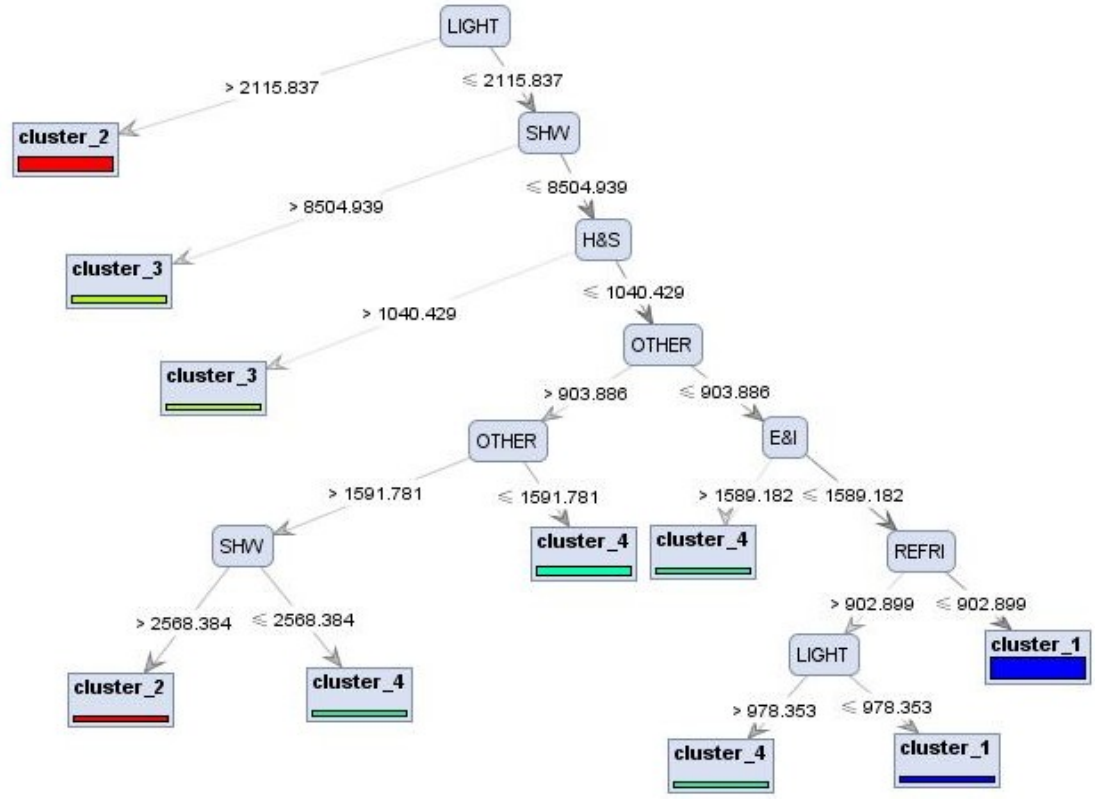

Figure 7-4 Decision tree for the prediction of cluster attribution

(2) Evaluation of the decision tree

Table 7-6 Confusion matrix

		Predicted data records			
		Cluster_1	Cluster_2	Cluster_3	Cluster_4
Actual data records	Cluster_1	7	0	0	0
	Cluster_2	1	4	0	0
	Cluster_3	1	0	1	0
	Cluster_4	2	0	0	4

In order to evaluate the accuracy of the generated decision trees, the RapidMiner analysis report also provides a confusion matrix for data analysts. In this study, a four-dimensional confusion matrix was built since the decision tree has four target variables, as shown in Table 7-6.

In this table, the rows indicate the number of actual data records used for testing in each cluster; and the columns represent the number of predicted data records generated by applying the decision tree to the actual data records. For example, the first column shows that 7 records in cluster_1 were correctly classified; while one record in cluster_2, one record in cluster_3, and two records in cluster_4 were misclassified into cluster_1. Therefore, the accuracy of this decision tree, which is also called '*recall*' in the data mining domain, can be calculated as (7+4+1+4)÷(7+4+1+4+1+1+2) = 80%, which is still acceptable despite the fact that it is relatively low. This may be partly ascribed to the small size of database. Moreover, data records in cluster_2, cluster_3, and cluster_4 are

misclassified into cluster_1 (at least one record in each cluster and four records totally), while data records in cluster_1 are not misclassified into the other clusters. Such information indicates that cluster_1 is more prone to be misclassified than the other clusters. This may have occurred since nearly half of the data records in the database are in cluster_1, which makes the decision tree more sensitive to this cluster. An even distribution among the four clusters in the database would possibly improve the accuracy. In addition, the sum of values in the matrix corresponds to the number of data records used for model testing. Clearly 20 records in the database were randomly selected by RapidMiner for testing, which also implies that 46 data records were used to establish the decision tree.

(3) Utilization of the decision tree

The decision tree can be utilized to predict the cluster attribution of new buildings according to the main end-use loads. Such predictions can be easily made by traversing a path from the root node to a leaf node. Take the node in the lower left corner in Figure 7-4 as an example. The prediction can be made as follows: for a building, if LIGHT $\leqslant$ 2115.837 and SHW $\leqslant$ 8504.939 and H&S $\leqslant$ 1040.429 and OTHER > 903.886 and OTHER > 1591.781 and SHW > 2568.384, then this building belongs to cluster_2.

Besides the prediction of cluster attribution, useful information can also be extracted from the decision tree so as to help understand building occupant behavior improvement. For example, various attributes are selected by the decision tree algorithm to split the

nodes; and their degrees of closeness to the root node determine the number of records impacted. Therefore, the closer an attribute is to the root node, the more significant it affects the cluster attribution. Clearly the attribute significance in the decision tree can be ranked as: LIGHT > SHW > H&S > OTHER > E&I > REFRI. Such information indicates a general descending order of occupant behavior deserving attention when modifying user activities in Japanese residential buildings. Moreover, among the seven end-use loads, KITCH does not appear in the decision tree. This may have occurred due to the narrow gap between energy consumption caused by KITCH-related occupant behavior among the four clusters (see Section 7.5.1), and thus KITCH has the weakest influence on the cluster attribution.

7.5.2 RB identification

In order to demonstrate the methodology, a *case building* with the most comprehensive household appliances was selected for case study. Table 7-7 shows the level 1 loads in this *case building*.

Table 7-7 End-use data in the *case building* (unit: MJ per capita per year)

SHW	LIGHT	KITCH	REFRI	E&I	H&S	OTHER	Sum
3882.699	582.052	250.600	1541.394	1799.530	621.743	336.592	9014.610

Based on the decision tree, the cluster attribution of the *case building* can be

predicted as follows:

Step 1 Examine the value of LIGHT, i.e., the attribute in the root node. Since LIGHT = 582.052, the node test in the right branch $LIGHT \leqslant 2115.837$ is satisfied, then go to the right-side child node;

Step 2 Examine the value of SHW. Since SHW = 3882.699, the node test in the right branch $SHW \leqslant 8504.939$ is satisfied, then go to the right-side child node;

Step 3 Examine the value of H&S. Since H&S = 621.743, the node test in the right branch $H\&S \leqslant 1040.429$ is satisfied, then go to the right-side child node;

Step 4 Examine the value of OTHER. Since OTHER = 336.592, the node test in the right branch $OTHER \leqslant 903.886$ is satisfied, then go to the right-side child node;

Step 5 Examine the value of E&I. Since E&I = 1799.530, the node test in the left branch $E\&I \leqslant 1589.182$ is satisfied, then go to the left-side child node, which is a leaf node. As a result, the decision tree in Figure 7-4 predicts that the case building belongs to cluster_4.

Comparing with the other three clusters, cluster_4, as shown in Table 7-4, can be characterized as the building group with high energy consumption in OTHER, medium high energy consumption in REFRI and E&I. Therefore, the *case building* occupants should manage to improve their behavior related to OTHER, REFRI, and E&I.

After the prediction of cluster attribution, the sum of the seven main end-use loads in the buildings in cluster_4 was calculated and ranked. Table 7-8 shows these loads and

their sum in the 14 buildings in cluster_4 in ascending order.

Table 7-8 The main end-use loads in the 14 buildings in cluster_4 (Unit: MJ per capita per year)

No.	SHW	LIGHT	KITCH	REFRI	E&I	H&S	OTHER	Sum
1	1691.656	744.428	1141.730	898.208	468.707	83.617	1670.297	6698.644
2	2757.408	981.880	662.657	645.977	388.737	317.828	1100.376	6854.487
3	1464.821	287.523	936.880	924.793	1958.911	504.171	845.352	6922.450
4	2471.123	865.524	1065.978	879.398	608.810	162.782	942.645	6996.259
5	1782.779	1099.852	322.597	1773.017	2092.484	142.018	556.186	7768.933
6	3337.796	558.252	411.807	1013.407	1060.430	360.339	1253.659	7995.690
7	3123.892	1094.065	1418.592	1055.741	803.612	160.549	1288.371	8944.821
8	2694.449	1758.554	621.970	1170.580	1109.116	503.125	1220.652	9078.446
9	3348.343	1407.656	1474.419	1046.065	768.032	550.396	739.591	9334.501
10	5224.677	617.440	724.771	565.889	498.162	186.758	1530.789	9348.487
11	4801.992	1080.952	994.315	909.184	870.845	202.665	818.539	9678.492
12	5192.053	982.723	768.211	777.985	363.490	923.699	1129.407	10137.568
13	5685.900	598.837	752.744	660.163	1007.248	269.102	1526.953	10500.947
14	2366.639	1089.153	451.300	2585.726	1878.995	817.197	2374.798	11563.808

A RB needs to be identified for the *case building* for the evaluation of energy-saving potential and the improvement of occupant behavior. The buildings with less total energy consumption (i.e., the sum of the seven main end-use loads) than the *case building* in cluster_4 were considered to be RB candidates. In order to provide reliable information for the *case building* occupants, the RB was defined as the most similar building to the *case building* in terms of building-related parameters. The Euclidean distance was used to determine the similarity. Various building-related parameters were captured from the

database to calculate the Euclidean distance, and among them, five are categorical parameters and are transformed into [0, 1], as shown in Table 7-9.

Table 7-9 Transformation of categorical parameters

Categorical parameters	CO		HT		Energy sources by usage (HEAT, HWS, KITC)	
Value	wood	non-wood	apartment	detached house	Electric	non-electric
Transformation value	0	1	0	1	0	1

Table 7-10 Building-related parameters of RB candidate buildings and the case building

No.	NO	FA	HLC	ELA	CO	HT	Energy sources by usage			T	V	RH	RA
							HEAT	HWS	KITC				
1	4	112	2.04	4.385	1	1	1	0	0	15.1	2.1	73	12.3
2	4	141.6	1.79	0.77	0	1	0	0	0	12.8	4.3	74	11.7
3	2	185.9	1.87	0.35	1	1	1	1	1	8.8	3.6	68	12.6
4	4	115	2.61	6.365	0	1	0	1	1	16.9	2.5	66	12.6
5	2	87.05	0.83	1.06	1	0	1	1	1	8.8	3.6	68	12.6
6	2	135	1.7	3.9	1	1	0	0	0	17.2	2.8	66	13.1
7	4	160.6	1.84	2.20	0	1	1	1	1	11.8	4.2	72	11.8
8*	2	128.3	1.69	0.6	0	1	0	1	1	8.8	3.6	68	12.6

* The *case building.*

Table 7-10 shows the building-related parameters of the RB candidate buildings and the *case building*.

Again, the min-max normalization was applied in order to help prevent attributes with large ranges from outweighing those with comparatively smaller ranges. After

normalization, the Euclidean distance between each candidate building and the *case building* was calculated; and the building with the smallest distance, i.e., No.3 building in Tables 7-10 and 7-8, was identified as the RB. For comparison, Table 7-11 shows the main end-use loads in the *case building* and the RB.

Table 7-11 Comparison of end-use data between the case building and RB (Unit: MJ per capita per year)

Building	SHW	LIGHT	KITCH	REFRI	E&I	H&S	OTHER	Sum
Case building	3882.699	582.052	250.6	1541.394	1799.53	621.743	336.592	9014.61
RB	1464.821	287.523	936.88	924.793	1958.911	504.171	845.352	6922.45

Table 7-11 shows that the sum of energy consumption in the *case building* is evidently higher than that in the RB. Further, user activities in the *case building* caused significantly higher energy consumption in SHW, LIGHT, REFRI, and H&S than that of the RB. This indicates that, in comparison with buildings with similar occupant behavior and building-related parameters, energy-saving potential still exists for the *case building*. That means energy consumption may be considerably reduced through modifying occupant behavior related to SHW, LIGHT, REFRI, and H&S. It should be noted that energy consumption in REFRI in cluster_4 is also medium high when comparing with the other three clusters. This implies the energy-saving potential of REFRI-related behavior is comparatively higher than the potential of the others, and thus deserves extra attention.

Additionally, energy-saving potential in the *case building* can be identified as the energy consumption difference between the two buildings, i.e., 9014.610 – 6922.450 = 2092.161 MJ per capita per year.

7.5. 3 Association rule mining (ARM)

Based on the information obtained from cluster-then-classification and RB identification, the ARM was then performed to find all the associations among the end-use loads at both levels. Accordingly, energy-inefficient specific occupant behavior will be determined and then energy-saving recommendations for modifying activities can be provided for the *case building* occupants.

After experimenting with various combinations of *support* and *confidence* values, a *support* of 50% and a *confidence* of 80% were set as minimum thresholds. Such thresholds mean that, for each generated association rule, at least 50% of all the data records under analysis contain both premise and conclusion; and the probability that a premise's emergence leads to a conclusion's occurrence is 80% or more. In addition, the minimum threshold of *lift* value was set 1 to find positive correlations. Such mining generated 756 rules, many of which are obvious and uninteresting; and truly interesting rules need to be further identified based on domain knowledge. Fifteen association rules between household appliances were selected for demonstration purposes, as shown in Table 7-12. It should be mentioned that most obtained associations are between attributes

in the LOW range (i.e., low energy consumption), while clearly the associations in the HIGH range (i.e., high energy consumption) may provide more useful information on energy conservation. This also indicates that the attributes involved in the obtained rules have a skewed distribution toward the LOW range, and may be ascribed to the high degree of building occupants' energy-saving consciousness. Moreover, due to the availability of the data source, daily data was used for ARM instead of hourly data; and thus the obtained rules do not necessarily indicate that user activities in the premises and conclusions occur simultaneously. Therefore, the actual occupant behavior patterns should also be taken into consideration when using these rules in practice.

The results of the cluster analysis show that the *case building* was grouped into cluster_4, which was characterized as the building group with high energy consumption in OTHER, medium high energy consumption in REFRI and E&I. Hence, association rules involving OTHER, REFRI and E&I are the most important and deserve more attention. Accordingly, two rules, i.e., *Rule 1* and *Rule 2* in Table 7-12, were found among all the obtained rules and discussed as follows:

Rule 1 shows that *living room outlet* and OTHER have a strong positive association with a *confidence* of 98% and a *lift* of 1.49. From this rule, it can be inferred that, in this building, the electricity load increase in *living room outlet* would quite possibly lead to the increase in OTHER. This indicates that, among all the unclear items included in OTHER, removable electrically-operated devices connecting to the living-room power

plugs deserve more attention than other devices. Therefore, building owners could easily identify these devices and then manage to modify their usage to reduce energy consumption.

Table 7-12 Selected association rules (*min_sup*[a*] = 50%, *min_conf*[b*] = 80%, *min_lift*[c*] =1)

No.	Premise	Conclusion	*Sup.*	*Conf.*	*Lift*
Rule 1	Living room outlet [LOW]	OTHER [LOW]	54%	98%	1.49
Rule 2	Heating boiler [HIGH]	REFRI [HIGH]	51%	94%	1.12
Rule 3	Lamp 1F [LOW]	LIGHT [LOW]	59%	96%	1.33
Rule 4	Washing machine 2F [LOW]	H&S [LOW]	76%	97%	1.25
Rule 5	Dishwasher [LOW]	KITCH [LOW]	74%	99%	1.26
Rule 6	Vacuum cleaner [LOW]	H&S [LOW]	67%	84%	1.07
Rule 7	Microwave, toaster, coffee [LOW]	KITCH [LOW]	66%	81%	1.04
Rule 8	TV (master bedroom 2F) [LOW]	Lamp 2F [LOW]	66%	87%	1.10
Rule 9	TV (other rooms) [LOW]	LIGHT [LOW]	51%	81%	1.11
Rule 10	Video [LOW]	Table lamp [LOW]	52%	84%	1.02
Rule 11	Lamp 1F [LOW]	Table lamp [LOW]	52%	84%	1.02
Rule 12	TV (Standby Power) [HIGH]	Ventilator [HIGH]	55%	100%	1.82
Rule 13	Phone [LOW]	Boom box [LOW]	57%	90%	1.06
Rule 14	TV (dining room) [LOW]	Boom box [LOW]	51%	85%	1.01
Rule 15	TV (other rooms) [LOW]	Boom box [LOW]	54%	86%	1.02

[a*] Minimum *support*, [b*] Minimum *confidence*, and [c*] Minimum *lift.*

Rule 2 shows that *heating boiler* has a strong positive association with REFRI with a *confidence* of 94% and a *lift* of 1.12. Given that the daily energy consumption of the heating boiler is mainly impacted by occupant presence and outdoor air temperature, this rule implies that, two factors (i.e., both a longer stay time of occupants and a lower

outdoor air temperature) possibly cause a higher energy consumption of refrigerators. With regard to the first factor, it sounds reasonable since a longer stay time of occupants tends to increase the refrigerator usage, thereby increasing the energy consumption. With regard to the second factor, it seems unreasonable since a low outdoor air temperature normally causes a relatively low indoor air temperature in a detached house without central HVAC systems, thereby decreasing the energy consumption of refrigerators. A possible explanation for this is that the building occupants had high thermal comfort requirements in cold days; and preferred to a high indoor air temperature by increasing the boiler thermostat setting or using kerosene space-heaters. In order to justify the assumption, the pattern relating mean daily kitchen air temperature[4] to mean daily outdoor air temperature was plotted, as shown in Figure 7-5. A trend line was then drawn to find out whether the kitchen air temperature increased or decreased in relation to outdoor air temperature. Clearly, a downward trend in mean daily kitchen air temperature following the increase of mean daily outdoor air temperature can be observed, which is in accordance with the assumption.

Therefore, a trade-off between human thermal comfort and building energy

[4]In this building, both the kitchen and the living room are in the first floor, and there are no partitions between them. Hence, they have the same indoor air temperature and the living room air temperature was used in this figure.

consumption is necessary for the owners, since an appropriate decrease of indoor thermostat settings for cold days results in an energy-consumption reduction in both space heating and refrigerators.

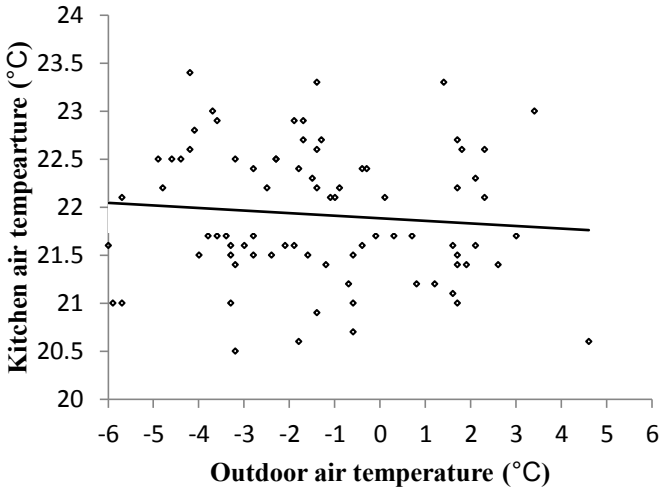

Figure 7-5 Mean daily air temperature in kitchen vs. mean daily outdoor air temperature (winter, 2003)

Further, the comparison between the RB and the *case building* shows that user activities in the *case building* caused significantly higher energy consumption in SHW, LIGHT, REFRI, and H&S than those in the RB. Hence, rules associating with these four attributes also deserve extra attention. At the same time, in order to provide more comprehensive recommendations for energy-efficient behavior, rules associating with other end-use loads were also analyzed in this study. Eventually, thirteen interesting rules

(i.e., *Rules 3* to *15* in Table 7-12) were selected and discussed as follows.

Similar to *Rule 1*, *Rules 3, 4,* and *5* show that *lamp 1F*, *washing machine 2F* and *dishwasher* have a strong positive association with LIGHT, H&S, and KITCH, respectively. *Rules 6* and *7* show that *vacuum cleaner,* and *microwave, toaster, coffee* have a positive association with H&S and KITCH, respectively. Therefore, comparing with other appliances associating with LIGHT, H&S, and KITCH, the building occupants should pay more attention to the use of lamps in the first floor, washing machines in the second floor, and dishwashers, since activities related to these appliances could have a major influence on the corresponding main end-use loads. At the same time, the use of vacuum cleaners, microwave ovens, toasters, and coffee machines also deserve some attention, though their associations with H&S and KITCH are weaker than *washing machine 2F* and *dishwasher*.

Rule 8 shows that *TV (master bedroom 2F)* has a positive association with *lamp 2F* with a *confidence* of 87% and a *lift* of 1.10. From this rule, it can be inferred that the usage of *TV (master bedroom 2F)* would quite possibly lead to the usage of *lamp 2F*. This may have occurred since the building occupants always turned the lights on when they were watching TV. An effective way of reducing energy consumption in this building is to watch TV with dim light.

Rules 9 to 11 can be explained in the same way as *Rule 8* and similar recommendations can be provided.

An unexpected result was that *TV (Standby Power)* and *Ventilator* have a strong positive association with a *confidence* of 100% and a *lift* of 1.82, as shown in *Rule 12*. Clearly the standby power of TVs and ventilators has the same trend of variation. This may have occurred since the building occupants would turn off the TVs and switch off the ventilators when the building was empty. However, standby power is commonly unnecessary and still accounts for energy cost. Therefore, TVs should be completely turned off or unplugged when they are not used. Furthermore, the wasted standby power of TVs is very small, but the sum of standby use consumed by all house appliances, such as microwave ovens, air conditioners, power adapters for laptop computers and other electronic devices, becomes significant. Standby power accounts for around 5-10% of residential electrical energy use in most developed countries; and continues to increase in developing countries (Standby Power, 2001). Hence, it is meaningful to help building owners to realize the importance of reducing standby power consumption, and feasible recommendations should also be provided for them. For example, a switchable power strip can be used for multiple devices, such as VCRs, DVD players, TVs, and computers, so that these appliances can be unplugged conveniently with one action.

Rules 13 to 15 show that *phone*, *TV (dining room)* and *TV (other rooms)* have a positive association with *boom box*. This indicates that, among all the appliances included in E&I, boom boxes was used in comparatively high frequency and deserve extra attention.

Moreover, indoor and outdoor parameters were also included in this ARM model. Associations between indoor/outdoor parameters and household appliances can assist in understanding the factors influencing occupant behavior. In order to demonstrate such associations, six rules were selected and shown in Table 7-13.

Table 7-13 Selected association rules between indoor/outdoor parameters and household appliances (*min_sup* = 50%, *min_conf* = 80%, *min_lift*=1)

No.	Premise	Conclusion	*Sup.*	*Conf.*	*Lift.*
Rule 1	Master bedroom (2F) temperature [HIGH]	Microwave, toaster, coffee [LOW]	58%	83%	1.02
Rule 2	Living room humidity [LOW]	Microwave, toaster, coffee [LOW]	55%	86%	1.06
Rule 3	Outdoor relative humidity [LOW]	Microwave, toaster, coffee [LOW]	57%	87%	1.07
Rule 4	Outdoor air temperature [LOW]	H&S [LOW]	54%	88%	1.12
Rule 5	Outdoor air velocity [LOW]	H&S [LOW]	59%	82%	1.05
Rule 6	Living room humidity [LOW]	H&S [LOW]	57%	90%	1.15

Rules 1 to *3* show that *master bedroom (2F) temperature (HIGH)*, *living room humidity*, and *outdoor relative humidity* have a positive association with *microwave, toaster and coffee*. This indicates that a high master bedroom temperature, as well as a low living room or outdoor relative humidity, tends to decrease the usage of microwave ovens, toasters, and coffee machines. A possible explanation for this is that the increase in indoor air temperature, or the decrease in indoor/outdoor relative humidity, causes the occupants to lose their appetite to some extent.

Rules 4 to *6* show that outdoor air temperature, outdoor air velocity, and living room humidity have a positive association with H&S. This indicates that the decrease in outdoor air temperature/velocity, and living room humidity tends to reduce the likelihood that occupants do housework such as cleaning and washing. It can be inferred that both local climatic conditions and indoor microclimate may have an impact on occupant behavior relating to housework. For example, the increase of outdoor air velocity may deteriorate indoor sanitary conditions (dust accumulation), thereby increasing the usage of vacuum cleaners and other sanitary appliances.

In addition, based on all the generated rules, it was found that six attributes, as shown in Table 7-14, have no association with the remaining attributes.

Table 7-14 Attributes without associations with the remaining attributes

No.	Appliances	Indoor parameters
1	Total energy consumption	Living room temperature
2	I&E	Bedroom (1F) temperature
3	Bidet	
4	IH heater	

The fact that these attributes have no association with the other attributes implies that, in this building, they are independent. There are two possible reasons for these attributes' independence: for *total energy consumption* and I&E, they may be decided by the holistic effects of various user activities, instead of associating with some certain activity. For the

other four attributes, their values may be purely random or remain relatively stable in the whole winter and thus no association with other attributes can be found. Such information can help building owners to make intelligent decisions when modifying their behavior.

7.6 Summary

A methodology for identifying and improving occupant behavior in existing residential buildings is developed. End-use loads of various household appliances were mapped onto corresponding occupant behavior, and were used to deduce user activities indirectly in this study. Specifically, these end-use loads were divided into two levels (main and sub-category), and thus correspond to two-level activities, i.e., general and specific occupant behavior.

In order to demonstrate its applicability, this methodology was applied to the residential buildings located in six different districts of Japan. A building with the most comprehensive household appliances was selected as the *case building* and the remaining buildings were used as related buildings. Data pre-processing was performed for the related buildings and they were grouped into four clusters by using K-means algorithm. The characteristic of occupant behavior in each cluster was analyzed. Base on these clusters, a decision tree was generated and its accuracy was evaluated as 80%. In terms of the decision tree, the *case building* was predicted to belong to cluster_4. A *reference building* was identified in the same cluster as the case building. Consequently, the *case building* was compared with buildings in the other clusters and the *reference building* to

determine energy-inefficient general behavior. Also, its energy-saving potential was identified as 2092.161 MJ per capita per year. Moreover, association rules were mined based on the data of the *case building* in winter in 2003, given the seasonality of occupant behavior. A number of interesting rules were found, and associations and correlations between different user activities were discovered. According to these rules, specific recommendations for highlighting energy-saving opportunities were provided for the building occupants.

The results obtained could help building occupants to modify their behavior, thereby significantly reducing building energy consumption. Moreover, given that the proposed method is partly based on the comparison with similar buildings, it could motivate building occupants to modify their behavior.

8. CONCLUSIONS AND RECOMMENDATIONS

8.1 Conclusions

In this dissertation, data mining are proposed to analyze measured building-related data. Furthermore, a data analysis process and a data mining framework are proposed to extract useful knowledge from building-related data, so as to help reduce building energy consumption. The process consists of eight steps: (1) problem definition and objective setting; (2) data source selection; (3) data collection; (4) data preprocessing/preparation; (5) data warehouses/marts construction; (6) data mining and model construction; (7) results analysis and evaluation; (8) knowledge discovery and presentation. The framework is composed of measured building-related data and data mining algorithms. It provides useful knowledge about the total building energy performance. In particular, three main data mining techniques, namely classification analysis, cluster analysis, and association rule mining are employed in this framework.

The applicability of the proposed process and framework was demonstrated through their applications to two sets of data collected from 80 residential buildings and a mechanically ventilated building. The applications have suggested that the process and framework can effectively help develop data analysis methodologies for extracting hidden useful knowledge from building-related data, in order to account for interactions between building energy consumption and its influencing factors. A clear and thorough

understanding of such interactions could provide essential guidance in reducing building energy consumption. In this study, four data analysis methodologies were developed and applied to the collected data, and are summarized as follows:

(1) Classification analysis was applied to develop a methodology for establishing building energy demand predictive models.. The developed model estimates the building energy performance indexes in a rapid and easy way. This methodology is appropriate to classify and predict categorical variables: its competitive advantage over other widely used modeling techniques, such as regression methods and ANN methods, lies in the ability to generate accurate predictive models with interpretable flowchart-like tree structures that enable users to quickly extract useful information. To demonstrate its applicability, the methodology was applied to estimate residential building energy performance indexes by modeling building energy use intensity (EUI) levels (either high or low). The results demonstrate that the decision tree method can classify and predict building energy demand levels with an accuracy of 93% for training data and 92% for test data, and identify and rank significant factors of building EUI automatically. The method can provide the combination of significant factors as well as the threshold values that will lead to high building energy performance. Moreover, the average EUI in each classified data subsets can be used as reference when performing prediction. The outcomes of this methodology could benefit architects, building designers and owners greatly in the building design and

operation stage. One crucial benefit is improving building energy performance mainly due to the fact that designers can optimize their building design plans based on the combination of significant factors as well as the threshold values. Another advantage of this methodology is that it can be utilized by users without requiring much computation knowledge.

(2) Cluster analysis was used to develop a methodology for examining the influences of occupant behavior on building energy consumption. To deal with data inconsistencies, min-max normalization is performed as a data preprocessing step before clustering. Grey relational grades, a measure of relevancy between two factors, are used as weighted coefficients of different attributes in cluster analysis. To demonstrate the applicability of the proposed methodology, it was applied to a set of residential buildings' measurement data. The results show that the methodology facilitates the evaluation of building energy-saving potential by improving the behavior of building occupants, and provides multifaceted insights into building energy end-use patterns associated with the occupant behavior. The results obtained could help occupants to prioritize efforts at the modification of their behavior in order to reduce building energy consumption.

(3) Association rule mining was employed to develop a methodology for examining all associations and correlations between building operational data, thereby discovering useful knowledge about energy conservation. To provide information for building

owners and operators to reduce energy consumption, both daily and annual data are mined. Moreover, data from two different years is mined, and the obtained associations and correlations in the two years are compared. In order to demonstrate the applicability of the proposed methodology, it was applied to the operational data of an air-conditioned building. The results show there are possibilities for saving energy by modifying the operation of mechanical ventilation systems and by repairing equipment. The results obtained from this methodology could help to better understand building operation and provide opportunities for energy conservation.

(4) Cluster analysis, classification analysis, and association rule mining were combined to formulate a methodology for identifying and improving occupant behavior in buildings. In order to demonstrate its applicability, the methodology was applied to a group of residential buildings, and one building with the most comprehensive household appliances was selected as the case building. The results show that, for the case building, the methodology was able to identify the behavior which needs to be modified, and provide occupants with feasible recommendations so that they can make required decisions to modify their behavior. Also, a reference building can be identified for the case building to evaluate its energy-saving potential due to occupant behavior modification. Considering the diversity of specific occupant behavior, the determination of energy-inefficient general occupant behavior can narrow down the scope of identification of energy-inefficient specific occupant behavior, and thus can

help occupants to quickly find the generated association rules, as well as specific behavior, which deserve more attention. Also, such information is extracted from the measured data and covers almost all energy-related behavior. With such information, building occupants can then better understand their behavior patterns, and easily focus on the energy-inefficient behaviour that needs to be modified. Therefore, the main advantage of the proposed methodology lies in its high efficiency of occupant behavior improvement. Moreover, the identification of energy-inefficient general behavior in this study is mainly based on the comparison with other similar buildings; this can help building owners to be aware of avoidable energy waste caused by their behavior, and motivate them to modify their activities accordingly.

8.2 Future Work

The recommendations for the future study include:

(1) The proposed data mining framework can be further improved in two aspects:

(i) Other data mining techniques also could be used besides the three data mining techniques employed in this framework. For example, anomaly detection (e.g., outlier and deviation detection) may be used in the field of fault detection and diagnosis. At the same time, summarization (e.g., visualization and report generation) may be used to help develop building automation systems.

(ii) Some data mining methods have been successfully used to address the problems within the building engineering domain in this research, such as the decision tree

method. However, more methods can be utilized to analyze measured building-related data and extract useful knowledge. For example, the regression tree method could be used to predict numerical variables in building energy demand modeling.

(2) The four proposed methodologies can be further improved as follows:

(i) With regard to the proposed methodology for modeling building energy demand, the main focus of future research should be placed on selecting appropriate interval number and reference value of target variables without reducing estimation accuracy, since these measures will provide more precise and valuable information to users. In addition, more case studies in different sectors, such as commercial buildings and office buildings, should be conducted to further benefit energy conservation and policy formulation.

(ii) With regard to the proposed methodology for identifying the effects of occupant behavior on building energy consumption, the main focus of future research should be placed on identifying appropriate building sample sizes and number of clusters, selecting typical attributes that can adequately represent the influencing factors unrelated to occupant behavior, since these measures will provide more precise effects of occupant behavior. In addition, more case studies in different sectors, such as commercial buildings and office buildings, should be conducted to further improve building energy performance and policy formulation.

(iii) The proposed methodology for examining all the associations and correlations between building operational data could be further improved by applying the proposed methodology to building operational data collected in different building sectors, climates, and building automation systems, in order to further evaluate its effectiveness and help understand the impact of different elements influencing building energy consumption. Once the methodology is generally accepted, it can be integrated into online data analysis and online fault detection to reduce building energy consumption efficiently. The software RapidMiner can be employed to perform the ARM and to help realize this methodology. Moreover, it can serve as a data mining engine for the integration and can automatically report interesting rules/patterns without requiring human intervention. However, data analysts are still necessary to compare obtained association rules to discover useful knowledge about building energy performance improvement.

(iv) The proposed methodology for identifying and improving occupant behaviour that needs to be modified in existing residential buildings could be further improved by identifying appropriate database sizes as well as the number of clusters, and improving the accuracy of the generated decision tree. The selection of database sizes and the number of clusters has a strong influence on grouping buildings and characterizing occupant behavior in terms of obtained building groups. The accuracy of the generated decision tree has a strong influence on assigning the *case*

building to obtained building groups. In addition, it is noted that using daily end-use loads in the *case building* to mine association rules and provide recommendations for occupants is not sufficient. This is because user activities in the premises and conclusions of association rules may not occur simultaneously. For example, user activities in the premises may occur in the morning while activities in the conclusions may occur in the afternoon. Consequently the recommendations made based on the analysis of association rules will be meaningless or even misleading. In order to overcome this limitation, hourly (or less than one hour, such as 15 minutes) end-use loads of various household appliances should be measured and used in association rule mining.

REFERENCES

AboulNaga, M.M., Elsheshtawy, Y.H. (2001). Environmental sustainability assessment of buildings in hot climates: the case of the UAE. *Renewable Energy, 24,* (3-4), 553-563.

Al-ajmi, F.F., Hanby, V.I. (2008). Simulation of energy consumption for Kuwaiti domestic buildings. *Energy and Buildings, 40,* (6), 1101-1109.

Anstett, M., Kreider, J.F. (1993). Application of neuronal network models to predict energy use. *ASHRAE Transactions 99. Part 1.*

Al-Mumin, A., Khattab, O., Sridhar, G. (2003). Occupants' behavior and activity patterns influencing the energy consumption in the Kuwaiti residences. *Energy and Buildings, 35,* (6), 549–559.

Andersson, S., Olofsson T., Ostin, R. (1996). Predictions of energy demand in buildings using neural network techniques on performance data. *Proceedings of the 4th Symposium on Building Physics in the Nordic Countries, Espoo, Finland.*

Aydinalp, M., Ugursal, V.I., Fung, A.S. (2002). Modelling of the appliance, lighting, and space cooling energy consumption in the residential sector using neural networks. *Applied Energy, 71,* (2), 87-110.

Bourgeois, D. (2005). Detailed occupancy prediction, occupancy-sensing control and advanced behavioral modeling within whole-building energy simulation. Ph.D. thesis, l'Universite Laval, Quebec.

Bouckaert, R.R. et al. (2009). *WEKA Manual for Version 3-7-0.* University of Waikato. New Zealand.

Balaras, C.A., Droutsa, K., Argiriou, A.A., Wittchen, K. (2002). Assessment of energy and natural resources conservation in office buildings using TOBUS. *Energy and Buildings 34,* (2), 135-153.

Balaras, C.A., Dascalaki, E., Gaglia, A., Droutsa, K. (2003). Energy conservation potential, HVAC installations and operational issues in Hellenic airports. *Energy and Buildings 35,* (11), 1105-1120.

Breiman, L., Friedman, J.H., Olshen, R.A., Stone, C.J. (1984). *Classification and Regression Trees,* Wadsworth, Inc., California.

Balaras, C.A., Gaglia, A.G. Georgopoulou, E., Mirasgedis, S., Sarafidis, Y., Lalas, D.P. (2007). European residential buildings and empirical assessment of the Hellenic building stock, energy consumption, emissions and potential energy savings. *Building and Environment, 42,* (3), 1298-1314.

Caldera, M., Corgnati, S.P., Filippi, M. (2008). Energy demand for space heating through a statistical approach: application to residential buildings. *Energy and Buildings, 40,* (10), 1972-1983.

Chow, T.T., Fong, K.F., He, W., Lin, Z., Chan, A.L.S. (2007). Performance evaluation of a PV ventilated window applying to office building of Hong Kong. *Energy and Buildings 39,* (6), 643-650.

Chung, W., Hui, Y.V. (2009). A study of energy efficiency of private office buildings in Hong Kong. *Energy and Buildings, 41,* (6), 696-701.

Chung, W., Hui., Y.V. (2009). A study of energy efficiency of private office buildings in Hong Kong. *Energy and Buildings, 41,* (6), 696-701.

Cabena, P., Hadjinian, P., Stadler, R., Verhees, J., Zanasi, A. (1998). *Discovering data mining: from concept to implementation*, Prentice Hall, Upper Saddle River, NJ.

Climate Statistics, Japan Meteorological Agency. Monthly Mean and Monthly Total Tables: http://www.data.jma.go.jp/obd/stats/data/en/smp/index.html

Crawley, D.B., Lawrie, L.K., Winkelmann, F.C., Buhl, W.F., Huang, Y.J., Pedersen, C.O., Strand, R.K., Liesen, R.J., Fisher, D.E., Witte, M.J., Glazer, J. (2001). EnergyPlus: creating a new-generation building energy simulation program. *Energy and Buildings 33,* (4), 319-331.

Cios, K.J., Pedrycz, W., Swiniarski, R.W. (2007). *Data mining: a knowledge discovery approach*, Springer, New York.

Catalina, T., Virgone, J., Blanco, E. (2008). Development and validation of regression models to predict monthly heating demand for residential buildings. *Energy and Buildings, 40,* (10), 1825-1832.

Cai, W.G., Wu, Y., Zhong, Y., Ren, H. (2009). China building energy consumption: Situation, challenges and corresponding measures. *Energy Policy, 37*, (6), 2054-2059.

Chen, S., Yoshino, H., Li, N. (2009). Statistical analyses on summer energy consumption characteristics of residential buildings in some cities of China. *Energy and Buildings, 42,* (1), 136-146.

Chen, S., Yoshino, H., Li, N. (2010). Statistical analyses on summer energy consumption characteristics of residential buildings in some cities of China. *Energy and Buildings, 42,* (1), 136-146.

Chen, S., Yoshino, H., Levine, M.D., Li, Z. (2009). Contrastive analyses on annual energy consumption characteristics and the influence mechanism between new and old residential buildings in Shanghai, China, by the statistical methods. *Energy and Buildings, 41,* (12), 1347-1359.

Cao, L.B., Yu, P.S., Zhang, C.Q., Zhang, H.F. (2009). *Data mining for business applications,* Springer, New York.

Deng, J.L. (1989). Introduction to grey system. *Journal of Grey System, 1,* 1-24.

Deng, S. (2003). Energy and water uses and their performance explanatory indicators in hotels in Hong Kong. *Energy and Buildings, 35,* (8), 775-784.

Davies, D.L., Bouldin, D.W. (1979). A Cluster Separation Measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *2,* 224.

Deng, S.M., Burnett, J. (2000). A study of energy performance of hotel buildings in Hong Kong. *Energy and Buildings, 31,* (1), 7-12.

Dong, B., Cao, C., Lee, S.E. (2005). Applying support vector machines to predict building energy consumption in tropical region. *Energy and Buildings 37,* (5), 545-553.

De la Flor, F.J.S. Lissén, J.M.S., Domínguez, S.Á. (2006). A new methodology towards determining building performance under modified outdoor conditions. *Building and Environment, 41,* (9), 1231-1238.

Dong, B., Lee, S.E., Sapar, M.H. (2005). A holistic utility bill analysis method for baselining whole commercial building energy consumption in Singapore. *Energy and Buildings, 37,* (2), 167-174.

Delgado, M., Sánchez, D.M., MartIn-Bautista, J., Vila, M.-A. (2001) Mining association rules with improved semantics in medical databases. *Artificial Intelligence in Medicine 21,* (1-3), 241-245.

Ekici, B.B., Aksoy, U.T. (2009). Prediction of building energy consumption by using artificial neural networks. *Advances in Engineering Software, 40,* (5), 356-362.

Emery, A.F., Kippenhan, C.J. (2006). A long-term study of residential home heating consumption and the effect of occupant behavior on homes in the Pacific Northwest constructed according to improved thermal standards. *Energy, 31,* (5), 677-693.

Eskin, N., Türkmen, H. (2008). Analysis of annual heating and cooling energy requirements for office buildings in different climates in Turkey. *Energy and Buildings 40,* (5), 763-773.

Frank, E., Hall, M., Trigg, L., Holmes, G., Witten, I.H. (2004). *Data Mining in Bioinformatics using Weka,* Bioinformatics Advance Access. Oxford University Press, England.

Filippín, C.S., Larsen, Flores. (2009). Analysis of energy consumption patterns in multi-family housing in a moderate cold climate. *Energy Policy, 37,* (9), 3489-3501.

Freire, R.Z., Oliveira, G.H.C., Mendes, N. (2008). Development of regression equations for predicting energy and hygrothermal performance of buildings. *Energy and Buildings 40,* (5), 810-820.

Fu, C., Zheng, J., Zhao, J., Xu, W. (2001). Application of grey relational analysis for corrosion failure of oil tubes. *Corrosion Science, 43,* (5), 881-889.

Gaunt, L. (1985). Habits and energy. *Meddelande M85:14,* The National Swedish Institute for Building Research, Sweden.

Ghiaus, C. (2006). Experimental estimation of building energy performance by robust regression. *Energy and Buildings, 38,* (6), 582-587.

Georgilakis, P.S., Gioulekas, A.T., Souflaris., A.T. (2007). A decision tree method for the selection of winding material in power transformers. *Journal of Materials Processing Technology, 181,* (1-3), 281-285.

Givoni, B., Kruger, E.L. (2003). An attempt to base prediction of indoor temperatures of occupied houses on their thermo-physical properties. *Proceeding of the Eighteenth International Passive and Low Energy Architecture Conference (PLEA'03), Santiago, Chile.*

Gaitani, N., Lehmann, C., Santamouris, M., Mihalakakou, G., Patargias, P. (2010). Using principal component and cluster analysis in the heating evaluation of the school building sector. *Applied Energy, 87,* (6), 2079-2086.

Hammarsten, S. (1979). A survey of Swedish buildings from the energy aspect. *Energy and Buildings, 2,* (2), 125-134.

Helsel D.R., (2002). *Hirsch R.M. Statistical methods in water resources.* U.S. department of the interior. U.S.

Hein, K.R.G. (2005). Future energy supply in Europe--challenge and chances. *Fuel, 84,* (10), 1189-1194.

Hsu, C.H. (2009). Data mining to improve industrial standards and enhance production and marketing: An empirical study in apparel industry. *Expert Systems with Applications 36,* (3), 4185-4191.

Hoes, P., Hensen, J.L.M., Loomans, M.G.L.C., De Vries, B., Bourgeois, D. (2009). User behavior in whole building simulation. *Energy and Buildings, 41,* (3), 295-302.

Han, J.W., Kamber, M. (2006). *Data Mining Concepts and Techniques 2nd ed.),* Elsevier Inc., San Francisco.

Hand, D., Mannila, H., Smyth, P. (2001). *Principles of data mining*, MIT press, Cambridge, MA.

HTML 4 Common Attributes: http://htmlhelp.com/reference/html40/attrs.html

Instruction for WEKA: http://weka.wikispaces.com/Primer

Jiménez, M.J., Heras, M.R. (2005). Application of multi-output ARX models for estimation of the U and g values of building components in outdoor testing. *Solar Energy 79,* (3), 302-310.

Jiao, J., Zhang, Y. (2005). Product portfolio identification based on association rule

mining. *Computer-Aided Design, 37,* (2), 149-172.

Kawashima, M. (1994). Artificial neural network back propagation model with three-phase annealing developed for the building energy predictor shootout. *ASHRAE Transactions 100, Part 2.*

Kreider, J.F., Claridge, D.E., Curtiss, P., Dodier, R., Haberl J.S., Krati, M. (1995). Building energy use prediction and system identification using recurrent neural networks, *Journal of Solar Energy Engineering, 117,* 161-166.

Krüger, E., Givoni, B. (2004). Predicting thermal performance in occupied dwellings. *Energy and Buildings, 36,* (3), 301-307.

Kim, Y.S., Kim, K.S. (2007). Simplified energy prediction method accounting for part-load performance of chiller. *Building and Environment 42,* (1), 507-515.

Kreider, J.F., Wang, X.A. (1992). Improved artificial neural networks for commercial building energy use prediction. *Proceedings of the ASME Annual Solar Engineering Meeting, Maui, HI.*

Kreider, J.F., Wang, X.A. (1997). Artificial neural network demonstration for automated generation of energy use predictors for commercial buildings. *ASHRAE Transactions 97, Part 2.*

Lam, J.C. (2000). Energy analysis of commercial buildings in subtropical climates. *Building and Environment, 35,* (1), 19-26.

Lazzarin, R.M., Castellotti, F., Busato, F. (2005). Experimental measurements and numerical modeling of a green roof. Energy and Buildings, 37, (12), 1260-1267.

Lopes, L., Hokoi, S., Miura, H., Shuhei, K. (2005). Energy efficiency and energy savings in Japanese residential buildings – research methodology and surveyed results, *Energy and Buildings, 37,* (7), 698-706.

Lam, J.C., Hui, S.C.M., Chan, A.L.S. 1997. Regression analysis of high-rise fully air-conditioned office buildings. *Energy and Buildings, 26,* (2), 189-197.

Li, Q., Meng, Q., Cai, J., Yoshino, H., Mochida, A. (2009). Applying support vector machine to predict hourly cooling load in the building. *Applied Energy, 86,* (10), 2249-2256.

Li, H., Nalim, R., Haldi, P.A. (2006). Thermal-economic optimization of a distributed multi-generation energy system--A case study of Beijing. *Applied Thermal Engineering 26,* (7), 709-719.

Lam, J.C., Wan, K.K.W., Cheung, K.L. (2009). An analysis of climatic influences on chiller plant electricity consumption. *Applied Energy, 86,* (6), 933-940.

Li, Y.M., Wu, J.Y., Shiochi, S. (2010). Experimental validation of the simulation module of the water-cooled variable refrigerant flow system under cooling operation. *Applied Energy , 87,* (5): 1513-1521.

Long, E.S., Zhou, J. (2005). Classified identifications: the annual relative variation rates (RVRs) of energy consumption are approximate in different cities with the same shading coefficient. *Building and Environment 40,* (4), 517-528.

Murakami, S., Akabayashi, S., Inoue, T., Yoshino, H., Hasegawa, K., Yuasa, K., Ikaga, T. (2006). Energy consumption for residential buildings in Japan, Architectural Institute of Japan, Maruzen Corp., http://tkkankyo.eng.niigata-u.ac.jp/HP/HP/database/index.htm.

Monts, J.K., Blissett, M. (1982). Assessing energy efficiency and energy conservation potential among commercial buildings: A statistical approach. *Energy, 7,* (10), 861-869.

MIT Technology review. (2001). Emerging Technologies That Will Change the World. http://www.technologyreview.com/Infotech/12265/

M. Aydinalp, V.I. Ugursal and A.S. Fung. Modeling of the space and domestic hot water energy consumption in the residential sector using neural networks. Applied Energy 2004; 79 (2): 159-178.

MacDonald, J.M., Wasserman, D.M. (1989). Investigation of metered data analysis methods for commercial and related buildings. *Report to U.S. Department of Energy under contract No. DE-AC05-84OR21400*, Oak Ridge National Laboratory, Oak Ridge, Tennessee.

Nakagami, H. (1996). Lifestyle change and energy use in Japan: Household equipment and energy consumption. *Energy, 21,* (12), 1157-1167.

Olofsson, T., Andersson, S. (2001). Long-term energy demand predictions based on short-term measured data. *Energy and Buildings, 33,* (2), 85-91.

Ourghi, R., Al-Anzi, A., Krarti, M. (2007). A simplified analysis method to predict the impact of shape on annual energy use for office buildings. *Energy Conversion and Management 48,* (1), 300-305.

O'Neill, P.J., Crawley, D.B., Schliesing, J.S. (1991). Using regression equations to determine the relative importance of inputs to energy simulation tools. *Proceedings of the Building Simulation '91 Conference Sophia-Antipolis, Nice, France, 1991.*

Ouyang, J., Hokao, K. (2009). Energy-saving potential by improving occupants' behavior in urban residential sector in Hangzhou City, China, *Energy and Buildings, 41,* (7), 711-720.

Ordenes, M., Marinoski, D.L., Braun, P., Rüther, R. (2007). The impact of building-integrated photovoltaics on the energy demand of multi-family dwellings in Brazil. *Energy and Buildings 39,* (6), 629-642.

Pan, H., Li, J., Zhang, W. (2007). Incorporating domain knowledge into medical image clustering. *Applied Mathematics and Computation, 185,* (2), 844-856.

Pérez-Lombard, L., Ortiz, J., Coronel, J. F., Maestre, I.R. (2011). A review of HVAC systems requirements in building energy regulations. *Energy and Buildings, 43,* (2-3), 255-268.

Pérez-Lombard, L., Ortiz, J., Pout, C. (2008). A review on buildings energy consumption information. *Energy and Buildings, 40,* (3), 394-398.

Priyadarsini, R., Wu, X.C., Lee, S.E. (2009). A study on energy performance of hotel buildings in Singapore. *Energy and Buildings, 41,* (12), 1319-1324.

Pan, Y., Yin, R., Huang, Z. (2008). Energy modeling of two office buildings with data center for green building design. *Energy and Buildings 40,* (7), 1145-1152.

Quinlan, J.R. (1986). Induction of decision trees. *Machine Learning, 1,* 81-106.

Quinlan, J.R. (1993). *C4.5 Programs for Machine Learning,* Morgan Kaufmann, San Mateo.

RapidMiner: http://rapid-i.com/content/view/181/190/

Reinhart, C.F. (2004). Lightswitch-2002: a model for manual and automated control of

electric lighting and blinds. *Solar Energy 77,* (1), 15-28.

Rokach, L., Maimon, O. (2008). *Data mining with decision trees: theory and applications,* SG: World Scientific, Singapore.

Rijal, H.B., Tuohy, P., Humphreys, M.A., Nicol, J.F., Samuel, A., Clarke, J. (2007). Using results from field surveys to predict the effect of open windows on thermal comfort and energy use in buildings. *Energy and Buildings 39,* (7), 823-836.

Shannon, C.E. (1948). A mathematical theory of communication. *The Bell System Technical J., 27,* 379-623.

Standby Power, Frequently Asked Questions (FAQs): http://standby.lbl.gov/faq.html

Santamouris, M., Balaras, C.A., Dascalaki, E., Argiriou, A., Gaglia, A. (1996). Energy conservation and retrofitting potential in Hellenic hotels. *Energy and Buildings, 24,* (1), 65-75.

Santamouris, M., Mihalakakou, G., Patargias, P., Gaitani, N., Sfakianaki, K., Papaglastra, M., Pavlou, C., Doukas, P., Primikiri, E., Geros, V., Assimakopoulos, M.N., Mitoula, R., Zerefos S. (2007). Using intelligent clustering techniques to classify the energy performance of school buildings. *Energy and Buildings, 39,* (1), 45-51.

Sullivan, R., Nozaki, S. (1984). Multiple regression techniques applied to fenestration effects on commercial building energy performance. *ASHRAE Transactions 90, Part 1A.*

Sullivan, R., Nozari, S., Johnson, R., Selkowitz, S. (1985). Commercial building energy performance analysis using multiple regression. *ASHRAE Transactions 91, Part 2A.*

Stevenson, W.J. (1994). Using artificial neural nets to predict building energy parameters, *ASHRAE Transactions 100, Part 2.*

Tian, Z., Love J.A. (2009). Energy performance optimization of radiant slab cooling using building simulation and field measurements. *Energy and Buildings, 41,* (3), 320-330.

Tonooka, Y., Liu, J., Kondou, Y., Ning, Y., Fukasawa, O. (2006). A survey on energy consumption in rural households in the fringes of Xian city. *Energy and Buildings, 38,* (11), 1335-1342.

Tso, G.K.F, Yau, K.K.W. (2007). Predicting electricity energy consumption: a comparison of regression analysis, decision tree and neural networks. *Energy, 32,* (9), 1761-1768.

Wu, S., Clements-Croome, D. (2007). Understanding the indoor environment through mining sensory data--A case study. *Energy and Buildings, 39,* (11), 1183-1191.

Westergren, K.-E., Högberg, H., Norlén, U. (1999). Monitoring energy consumption in single-family houses. *Energy and Buildings, 29,* (3), 247-257.

Yu, P.C.H., Chow, W.K. (2001). Energy use in commercial buildings in Hong Kong *Applied Energy, 69,* (4), 243-255.

Yu, Z., Fung, B.C.M., Haghighat, F., Yoshino, H., Morofsky., E. (2011). A systematic procedure to study the influence of occupant behavior on building energy consumption. *Energy and Buildings, 43,* (6), 1409-1417.

Yu, Z., Haghighat, F., Fung, B.C.M., Morofsky., E. (2011). A Methodology for identifying and improving occupant behavior in residential buildings. *Energy, 36,* (11), 6596-6608.

Yu, Z., Haghighat, F., Fung, B.C.M., Yoshino, H. (2010). A decision tree method for building energy demand modeling, *Energy and Buildings, 42,* (10), 1637-1646.

Yu, Z., Haghighat, F., Fung, B.C.M., Zhou, L. (2012). A novel methodology for knowledge discovery through mining all associations between building operational data. *Energy and Buildings,* In press.

Yao, Y., Lian, Z., Hou, Z., Liu, W. (2006). An innovative air-conditioning load forecasting model based on RBF neural network and combined residual error correction. *International Journal of Refrigeration, 29,* (4), 528-538.

Yang, J., Rivard, H., Zmeureanu, R. (2005). On-line building energy prediction using adaptive artificial neural networks. *Energy and Buildings, 37,* (12), 1250-1259.

Zhang, Q. (2004). Residential energy consumption in China and its comparison with Japan, Canada, and USA. *Energy and Buildings 36,* (12),1217-1225.

Zmeureanu, R., Fazio, P. (1991). Analysis of the energy performance of office buildings in Montreal in 1988. *Energy and Buildings, 17,* (1), 63-74.

Zmeureanu R., Fazio P., DePani S., Calla R. (1999). Development of an energy rating system for existing houses. *Energy and Buildings, 29,* (2), 107-119.

Zhou, Y.P., Wu, J.Y., Wang, R.Z., Shiochi, S., Li, Y.M. (2008). Simulation and experimental validation of the variable-refrigerant-volume (VRV) air-conditioning system in EnergyPlus. *Energy and Buildings, 40,* (6), 1041-1047.