Detecting Anomalies in Activity Patterns of Lone Occupants from Electricity Consumption Data

by

Kuan Long Leong

B.A.Sc., The University of British Columbia, 2013

A THESIS SUBMITTED IN PARTIAL FULFILMENT OF
THE REQUIREMENTS FOR THE DEGREE OF

MASTER OF APPLIED SCIENCE

in

THE FACULTY OF GRADUATE AND POSTDOCTORAL STUDIES

(Electrical and Computer Engineering)

THE UNIVERSITY OF BRITISH COLUMBIA

(Vancouver)

January 2016

Abstract

As the global population is ageing, the demand for elderly care facilities and services is expected to increase. Assisted living technologies for detecting medical emergencies and assessing the wellness of the elderly are becoming more popular. A person normally performs activities of daily living (ADLs) on a regular basis. A person who is able to perform recurring ADLs indicates a certain wellness level. Anomalies in activity patterns of a person might indicate changes in the person's wellness. A method is proposed in this thesis for detecting anomalies in activity patterns of a lone occupant using electricity consumption measurements of his/her residence. The proposed method infers anomalies in activity patterns of an occupant from electricity consumption patterns without a need of explicitly monitoring the underlying individual activities. The proposed method provides a score which is a quantitative assessment of anomalies in the electricity consumption pattern of an occupant for a given day. A survey was conducted to obtain the hourly activities of three lone occupants for a month. The level of suspicion values, which are quantitative assessments of anomalies in the daily activity patterns of the occupants, were deduced from the survey. Using Fuzzy C-Means (FCM) clustering with Euclidean distance measure, the scores and level of suspicion values were clustered respectively. A

day was then classified as regular or irregular based on the clustering results of the scores and level of suspicion values respectively. The results showed that anomalies in electricity consumption patterns can effectively reflect anomalies in the underlying activity patterns. The results also showed that the proposed feature and model based method outperforms a chosen raw data based approach. The performance of the proposed method was improved when subsets of features were considered based on the minimum Redundancy Maximum Relevance (mRMR) feature selection. A supervised learning method based on the Curious Extreme Learning Machine (C-ELM) was then proposed. The proposed method based on C-ELM (PM-CELM) outperforms the proposed method based on FCM (PM-FCM), but PM-FCM can operate without labelled training data.

Preface

This thesis is original, unpublished, independent work by the author, Kuan Long Leong, under the supervision of Professor Cyril Leung.

Table of Contents

List of Tables

List of Figures

List of Acronyms

ADL	Activity of Daily Living
C-ELM	Curious Extreme Learning Machine
CGH	Comparative Genomic Hybridization
FCM	Fuzzy C-Means Clustering
MID	Mutual Information Difference
MIQ	Mutual Information Quotient
mRMR	minimum Redundancy Maximum Relevance Feature Selection
PM-CELM	Proposed Method based on C-ELM
PM-FCM	Proposed Method based on FCM
PM-mRMRnCELM	Proposed Method based on mRMR and C-ELM
PM-mRMRnFCM	Proposed Method based on mRMR and FCM

SVM-RFE	Support Vector Machine Recursive Feature Elimination

Acknowledgements

I would like to express my sincere gratitude to my supervisor, Professor Cyril Leung, for his immeasurable support and guidance throughout my graduate studies. His patience and guidance help me overcome many challenges and finish this thesis. Without Professor Leung's guidance, this thesis would not have been possible.

This work was supported in part by the Natural Sciences and Engineering Research Council (NSERC) of Canada under Grant RGPIN 1731-2013, by the UBC Faculty of Applied Science, and by the National Research Foundation, Prime Minister's Office, Singapore under its IDM Futures Funding Initiative and administered by the Interactive and Digital Media Programme Office.

I would like to thank my co-supervisor, Professor Chunyan Miao (Nanyang Technological University), for her guidance and assistance in this research work. I would also like to thank Dr. Qiong Wu (Nanyang Technological University) for helping me understand the Curious Extreme Learning Machine (C-ELM).

I am thankful to Dr. Christine Chen for her suggestions on the topics for future research.

I am grateful to my family members and friends for providing me with the data

required for this research work and participating in the survey. Without their help, it would not have been possible to validate my research ideas.

My friends have helped me stay sane through the difficult years of graduate school. I greatly cherish their friendships and deeply appreciate their support.

Most importantly, none of this would have been possible without the love, encouragement and patience of my parents. I am deeply grateful to my parents who encourage me to study abroad and to pursue further education. Without my parents, I would not be who I am today.

I dedicate this thesis to my beloved parents.

Chapter 1

Introduction

This chapter begins with the motivation of the research work in this thesis. Works related to assisted living technologies and time series clustering are then discussed, followed by an outline of the contributions of this work. The organization of the thesis is described at the end of this chapter.

1.1 Motivation

The global population share of people aged 60 and over is expected to increase from 11.7% in 2013 to 21.1% in 2050 [1]. In other words, the number of the elderly people is expected to reach 2 billion in 2050 from 841 million in 2013. About 92% of people aged 65 and over in the United States have at least one chronic disease [2]. Similarly, almost 90% of people aged 65 and over in Canada have at least one chronic condition [3, 4] and 74% of people aged 65 and over in Canada are taking at least one medication [5]. The demand for elderly care facilities and services such as nursing homes is expected to increase with the ageing population. A U.S. survey found that 87% of people aged 65 and over prefer to age in their own homes [6]. Thus, there is a need to ensure a safe and independent

ageing environment for those elderly people who prefer to stay in their own homes, and to help them live in their preferred environments for as long as possible.

Assisted living technologies for detecting medical emergencies and assessing the wellness of the elderly are becoming more popular [7]. While the most direct way of detecting medical emergencies is arguably by monitoring physiological data such as heart rate or blood pressure, estimating the wellness of a person usually involves monitoring the activities of daily living (ADLs). An important component of the assisted living technologies is activity recognition and monitoring [7].

A person normally performs ADLs on a regular basis. The capability of performing ADLs regularly implies that the person is at least physically able to maintain a regular lifestyle. It also indicates that the wellness of the person is at a certain level. Large deviations from a regular daily routine may indicate changes in the capability of performing ADLs. Such deviations may be used to alert relatives or caregivers to look into the cause of the deviations in a timely manner.

To establish the regular activity patterns of a person, number of approaches have been proposed to monitor the individual ADLs [7]. Individual activities are usually monitored using ambient sensors (e.g. motion sensors and force sensors) placed smartly in the monitored environment or by cameras. Individual activities are then recognized as sequences of sensor events or sequences of images. The regular activity patterns of a person can then be established based on the sequences of recognized activities.

ADLs usually involve using electric home appliances. Activities that consume energy

might be inferred from the household energy consumption patterns. *Energy* denotes electrical energy or electricity in this section and the rest of the thesis unless otherwise specified. If a person's activity patterns are exactly the same every day, the energy consumption patterns will also probably be the same every day. The energy consumption patterns of a monitored environment would reflect the activity patterns of the person. If the activity patterns of a person can be sufficiently represented by the energy patterns, these could be used to detect anomalies in the underlying activities and there might be no need to explicitly monitor the underlying activities individually. This thesis aims to show that deviations from a person's regular activity patterns can be effectively detected using energy consumption measurements of his/her residence.

1.2 Related Works

Various assisted living technologies have been proposed to alleviate health problems, detect medical emergencies and improve the wellness of the elderly [8–30]. One important assisted living technology is health status monitoring. Physiological parameters are possibly the best indicators of the health status of a person. The challenge of developing technologies for monitoring the health status of a person in the home environment is to develop a portable, power-efficient and cost-effective device which is able to communicate with the relevant health care providers in a timely manner. Practical approaches proposed in [8–11] can monitor the physiological parameters such as electrocardiogram signals, blood glucose concentrations, blood pressure, body temperature, heart rate and

respiratory rate using ZigBee or Bluetooth for communication.

Another important assisted living technology is fall detection. The challenge in detecting falls in the home environment is to differentiate unintentional falls from normal activities and to minimize false alarms and missed detections. Fall detection approaches can be divided into two categories: wearable and non-wearable. Approaches using wearable accelerometers were proposed in [12, 13]. How the wearable device is worn and the person's willingness to wear the device are critical to the effectiveness of this approach. Audio based [14, 15] and visual based [16, 17] non-wearable approaches have also been proposed. Audio based approaches can be adversely affected by background noise while visual based approaches can be harmfully affected by occlusions.

Another important assisted living technology is activity recognition and monitoring. Activity recognition helps determine what daily activities are essential to a person. Long term monitoring helps determine the regular activity patterns of a person. A person who is able to perform recurring ADLs indicates a certain wellness level. Activity recognition approaches can generally be categorized as video-based [18] or sensor-based [19–30]. In [21–25], the authors designed a wireless sensor network which can interpret the wellness of a person by monitoring various home accessories such as bed, microwave oven, toilet, dining chair, etc. The wellness of a person was interpreted according to how well the monitored person performed the essential daily activities in terms of the home appliances' active and inactive durations. The difficulties of this approach are as follows: 1) determining a sufficient number of sensors for monitoring ADLs, 2) storing the sensor

data efficiently, 3) annotating the activities deduced from the sensor data, and 4) classifying regular and irregular activities accurately. Although an irregular activity pattern of a person might be detected correctly, the irregularity may or may not indicate a change in the person's wellness.

In [26], the authors tried to relate a person's movement pattern to physical and mental health. They explored the correlations of the movement patterns of 10 lone occupants in 10 different homes and baseline measures of their depression levels and mobility levels. The movement patterns were captured by motion sensors placed in the monitored homes. Although the movement patterns showed strong correlation with the baseline mobility levels as expected, they only show weak correlation with the baseline depression levels. The authors concluded that there was not sufficient evidence to show a significant correlation between one's movement patterns captured by the sensors and one's depression level. No follow-up study has been found by the author.

In [27], the authors proposed to detect anomalies in activity patterns of a person based on temporal relations between events captured by some sensors such as motion sensors and light sensors. For instance, if event A always occurs before event B but the recurring temporal relation between A and B is violated on a particular day, it will be noted as an anomaly. It would be impractical to investigate all events that occur during a day and hence the authors only focused on the temporal relations between the most frequent events. However, annotating the events captured by the sensors and identifying the temporal relations between events could still be time-consuming.

This work proposes to infer anomalies in activity patterns of a person from the household energy consumption patterns. Energy consumption data is essentially a time series or a sequence. A straightforward way would be to compare the energy sequences with one another using some similarity measure. Similar sequences are assigned to the same cluster and there could be multiple clusters. Each cluster can then be assigned a class label (e.g. regular or irregular) if there is sufficient information. Approaches to cluster time series data can be divided into three categories: 1) raw data based, 2) feature based, and 3) model based [31]. In [32], the authors compared four different commonly used similarity measures for clustering (Euclidean distance, Mahalanobis distance [33], Dynamic Time Warping distance [34] and Pearson's correlation [35]) by using raw hourly energy consumption data for five university buildings. It was found that the Euclidean distance measure was the overall best similarity measure for clustering the raw hourly energy consumption data according to four different validity techniques (Dunn index [36], Davies-Bouldin index [37], clustering balance [38] and cluster-vector balance [32]).

However, most similarity measures are too sensitive to slight changes in the raw hourly energy consumption data. For instance, two 24-hour hourly energy consumption sequences can be clustered into two different groups due to trivial differences even though the underlying activity patterns are almost the same. It has also been observed that some key features (e.g. maximum, minimum) of energy patterns are more relevant to the underlying activity patterns. In other words, the key features could be better indicators of the underlying activity patterns than the raw energy data. Therefore, it might be

better to work with features selected or extracted from the raw energy consumption data for detecting anomalies in activity patterns.

1.3 Contributions

This work explores the correlation between the household energy consumption data and the activity patterns of lone occupants living at home. A method is proposed to detect anomalies in activity patterns of the occupant by monitoring the household energy consumption. This work differs from related works in one fundamental perspective. This work intends to detect anomalies in activity patterns of a person without explicitly monitoring the individual activities or actions of the person. The household energy consumption patterns are used as representations of the activity patterns of the occupant without explicitly considering the individual activities. Instead of attempting to use the raw energy consumption data for detection, we use features extracted from the raw energy consumption data. The extracted features are designed to effectively reflect the underlying activities in terms of the time of day and related energy consumption. We show that the proposed method is more accurate for detecting anomalies in activity patterns of the occupant than a chosen raw data based approach.

If the objective of a certain system is to detect anomalies in activity patterns of a person, the proposed method represents a simple solution since annotating the activity data is generally time-consuming and usually requires a large training set [7]. If there is a strong correlation between the household energy consumption patterns and the activity

patterns of the occupant, the anomalies in energy consumption patterns will reflect the anomalies in activity patterns of the occupant. The method proposed in this work does not identify whether the anomalies detected indicate a positive or negative change in the person's wellness. However, it can trigger an alert to caregivers or relatives who can then investigate in a timely manner.

1.4 Structure of the Thesis

The rest of the thesis is organized as follows. In Chapter 2, a method for detecting anomalies in activity patterns of lone occupants from household energy consumption data is proposed. The datasets are first described, followed by the features used for representing regular energy patterns of an occupant and the model for quantitatively assessing the detected anomalies in energy patterns. In Chapter 3, the survey of activities and regular activity patterns are first described, followed by configurations of the relevant variables and thresholds. The correlation between energy patterns and activity patterns and a comparison of the proposed method with a chosen raw data based approach are then discussed. In Chapter 4, an overview of feature selection methods is first presented, followed by an introduction to the mRMR feature selection [39, 40] and a performance evaluation of the proposed method with reduced feature sets. In Chapter 5, a supervised learning method based on C-ELM [41] for classifying the energy patterns is proposed. The main findings and some topics for future research are summarized in Chapter 6.

Chapter 2

Detecting Anomalies in Activity Patterns from Electricity Consumption Data

The objective behind the proposed method is to detect anomalies in activities of daily living (ADLs) of an occupant from the household electricity (electrical energy) consumption data. A quantitative assessment of anomalies detected during a day is provided by the method. The result can be used to classify the daily activity pattern of the occupant as regular or irregular from an electrical energy consumption perspective. The method is designed for use in homes with lone occupants. The datasets, features for representing regular electrical energy consumption patterns, and model for quantitatively assessing detected anomalies are discussed in this chapter.

2.1 Datasets

The proposed method is based on hourly energy consumption data of a household over multiple days (e.g. 30 days). Data used in this work were collected from three participants living alone at home. As the BC Hydro smart meters [42] were installed in the participants' homes, the hourly electrical energy consumption data were simply downloaded by each participant from the BC Hydro website . A summary of the three datasets is shown in Table 2.1.

Home	# of Occupant	Type of Home	Data Length (days)
A	1	Apartment	30
B	1	Apartment	365
C	1	Detached	365

Table 2.1: Summary of the energy datasets

Most days of the dataset were 24-hour except for the first day and the last day of the Daylight Saving Time period. The 25-hour day was converted to a 24-hour day by assigning the average of the hourly consumptions of the two 1AM-2AM periods to one single 1AM-2AM period. The 23-hour day was converted to a 24-hour day by assigning the average of the hourly consumptions of 1AM-2AM and 3AM-4AM to the missing 2AM-3AM.

2.2 Features for Representing Regular Electrical Energy Consumption Patterns

Features are needed to represent the regular (recurring) energy consumption patterns of an occupant. The basic assumption made in this thesis is that people perform their ADLs on a regular basis. Many of the ADLs involve using electric home appliances. It is therefore plausible that these activities may be inferred from the electricity consumption data.

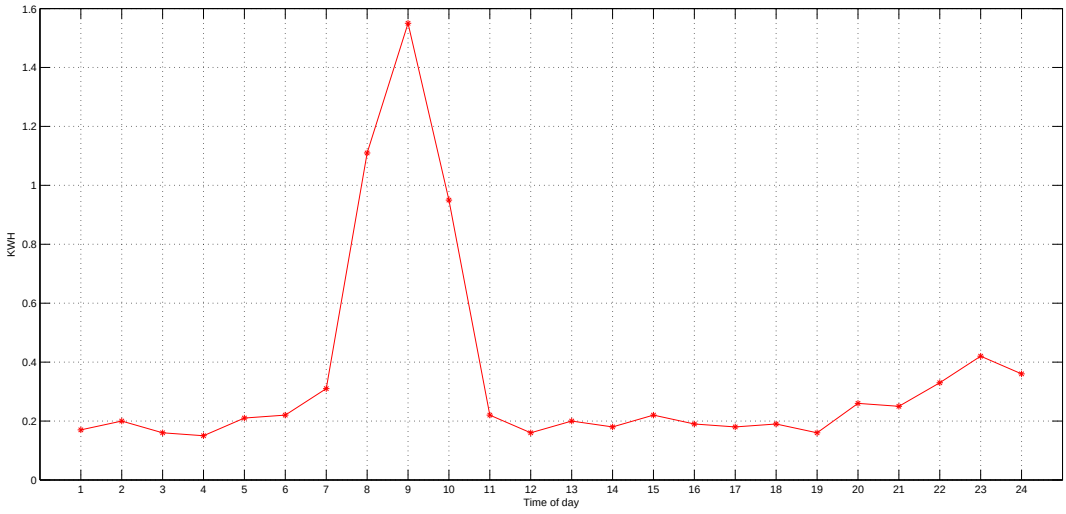

Figure 2.1: The typical hourly energy consumption pattern of Home C during a day

For instance, the occupant of Home C normally wakes up at 8AM, uses electric cooking appliances right away for two to three hours, is away from home from 11AM to 9PM, and goes to bed at 11PM. The hourly energy consumption pattern of a typical day at Home C is shown in Figure 2.1. As can be seen from Figure 2.1, the energy consumption pattern matches quite well the activity pattern. This provides strong evidence that the activity

patterns of the occupant are reflected in the energy consumption patterns. Instead of attempting to recognize the underlying individual activities from the energy consumption data, the proposed method uses energy patterns to represent activity patterns of the occupant without explicitly considering the underlying individual activities.

The available dataset in this work can only provide up to one data point per hour. However, an activity is not limited to begin and end within a single pre-framed hourly interval in the dataset. An activity can begin at any time in one hourly interval and end at any time in another hourly interval. For instance, if a certain activity normally takes one hour and it normally takes place within two hourly frames, the energy consumption corresponding to this activity can be distributed differently among the two consecutive hourly frames depending on the time at which the activity actually begins.

For example, if there is unrealistically only one given activity that can occur within two particular consecutive hourly frames and it consumes the exact same amount of energy whenever it occurs, the sum of the energy consumptions of the two hourly frames will always be the same whenever the activity occurs. However, if the two consecutive hourly frames are observed separately, the electricity consumptions in the first hour and in the second hour of the two hourly frames may not be the same respectively every day depending on when the activity occurs. Although there is a consistency in the activity pattern, it indicates that a consistency in this case may not be found in the energy consumption pattern if the energy consumptions of the two consecutive hourly frames are considered separately. However, a consistency can be found in the energy consumption

pattern if the sum of the energy consumptions of the two consecutive hourly frames is considered.

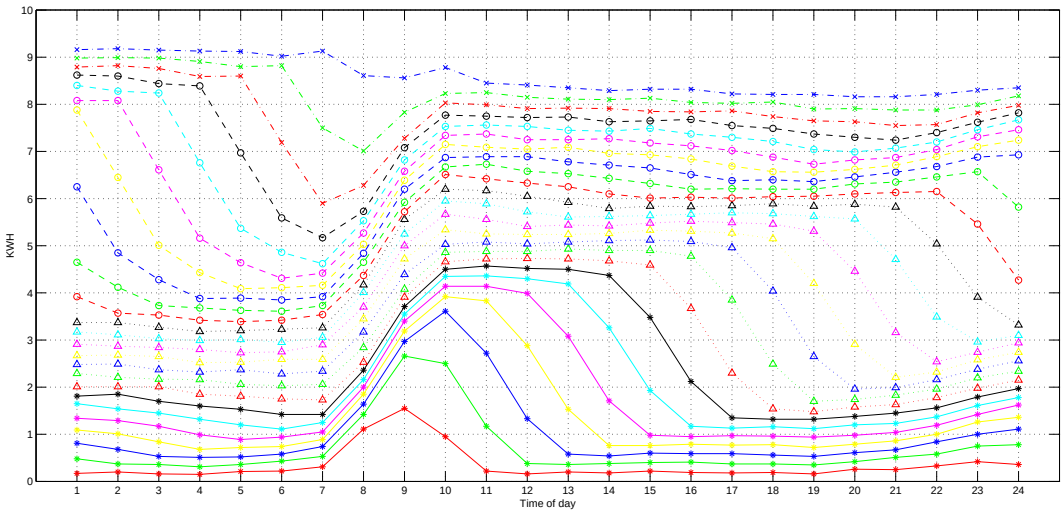

Figure 2.2: The 1-hour to 24-hour moving totals at each time point (hour)

It can easily be seen that a similar argument can also be made for any multiple consecutive hourly frames. This suggests that considering the multiple-hour total consumption can help reflect the underlying activity patterns of the occupant. Therefore, at each time point, the proposed method goes back in time to calculate the 1-hour to 24-hour total energy consumptions (moving totals). Consequently, at each time point, we not only have the original energy consumption of the previous hour but also the total energy consumptions of the previous two to 24 hours respectively, as illustrated in Figure 2.2. In Figure 2.2, at any time point, each dot on the bottom curve represents the hourly consumption of the previous one hour, each dot on the second curve from the bottom represents the total consumption of the previous two hours, and so on. This

step of the proposed method does not add any new information to the dataset; it simply computes the two to 24 hours moving totals using the hourly energy consumption data in the dataset.

The objective of the proposed method is to detect anomalies in activity patterns from the household energy consumption data. However, the raw energy consumption patterns might not be the best representations of the underlying activity patterns. This is because unimportant parts of the raw energy patterns might adversely affect the detection of anomalies in the underlying activity patterns. Although the activity patterns are the same, their corresponding energy patterns may not be exactly the same due to some small differences. Detecting anomalies may be ineffective if we compare one raw energy pattern to another using some conventional distance measure such as Euclidean distance measure. Rather, we should choose features which are more representative of the underlying activities from the raw energy patterns.

The local maxima and local minima of energy patterns are likely to be good indicators of the active and inactive hours of an occupant. Therefore, the times at which the maxima and minima occur may be good indicators of the underlying activity patterns. The difference between the maximum and minimum energy consumptions might indicate the energy consumption of the relevant activities. Therefore, the features chosen to represent the energy patterns of an occupant are: 1) the time of day when the maximum occurs for each moving total, 2) the time of day when the minimum occurs for each moving total, and 3) the range (difference between the maximum and minimum) of each moving total.

In the case that the extremum (maximum or minimum) is not unique, one time point among the multiple time points sharing the same extremum is chosen uniformly at random. Eventually, the features chosen to represent the energy consumption pattern of an occupant on a particular day are the 24 time points of maxima, 24 time points of minima, and the 24 ranges. One set of the 72 features obtained from one single day only represents the energy pattern of the occupant on that particular day. To deduce the regular energy patterns of the occupant, these features need to be collected over multiple days.

2.3 Model for Quantitatively Assessing Detected Anomalies

The model is used to detect anomalies in energy consumption patterns of an occupant and to provide quantitative assessments of the detected anomalies. The result can then be used to classify the daily energy consumption pattern of the occupant as either regular or irregular. Regular (recurring) energy patterns of the occupant are deduced from a collection of the above-mentioned features over multiple days. The regular electrical energy patterns, final score and design variables, selection of the electrical energy consumption data, probability thresholds and score assignment, z-score threshold and score assignment, and flexibility are discussed in this section.

2.3.1 Regular Electrical Energy Patterns

As mentioned in Section 2.2, the energy pattern of an occupant on a particular day is represented by the 24 time points of maxima, 24 time points of minima, and the 24 ranges. A quantitative score is provided by the proposed method to indicate how well an energy pattern of an occupant on a particular day matches the regular energy patterns. Regular energy patterns of an occupant are deduced from a collection of the features mentioned in Section 2.2 over multiple days (e.g. 60 days). The regular energy patterns of an occupant are described by 1) how likely a maximum or minimum of a moving total occurs at a particular time, and 2) how likely the range (maximum minus minimum) of a moving total happens to have a certain value. The time of day when a maximum or minimum energy consumption normally occurs reflects the active and inactive hours of an occupant. The normal range of a moving total indicates the typical total energy consumption of the relevant activities.

Using the time points of maxima and the time points of minima of each day over multiple days (e.g. 60 days), the probability of each time point (hour) being the maximum or minimum of each moving total can be approximated. The pseudocode of the algorithm for estimating the occurrence probabilities of the maximum (minimum) at each hour for the hourly energy consumption data is given in Algorithm 1.

Algorithm 1 The pseudocode for estimating the occurrence probabilities of the maxima (minima) for the hourly energy consumption data

1: **Step 1** Obtain the hourly energy consumption of each hour for one day (i.e. the bottom curve in Figure 2.2)

2: **Step 2** Record the time point at which the maximum (minimum) occurs

3: **Step 3** Repeat **Step 1** and **Step 2** for each day over multiple days (e.g. 60 days)

4: **Step 4** Estimate the probability that the maximum (minimum) would occur at 1AM according to the records

5: **Step 5** Repeat **Step 4** for the remaining 23 time points

To calculate the occurrence probability of the maximum (minimum) at each time point (hour) for the 2-hour moving total, Algorithm 1 obtains the *2-hour total energy consumption* of each hour for one day (i.e. the second to bottom curve in Figure 2.2) in **Step 1**. **Step 2** to **Step 5** are then performed. Corresponding modifications to **Step 1** of Algorithm 1 can be made in order to estimate the occurrence probabilities of the maxima (minima) for the 3-hour to 24-hour moving totals.

Following the previously described procedure (Algorithm 1 and its modified versions), a total of $24 \times 24 = 576$ occurrence probabilities are obtained for the maxima and an equal number for the minima. These probabilities can be arranged in two 24×24 matrices, one for the maxima and one for the minima. Part of a sample maxima probability matrix is illustrated in Figure 2.3, where each row represents a moving total and each column represents a time point. For instance, it shows that there is 63% chance that the

maximum of the 2-hour moving total occurs at 9AM. The minima probability matrix is not shown here, but it can be interpreted in a similar manner. These two matrices help represent the regular energy consumption patterns of an occupant and they can indicate the normal active and inactive hours of the occupant.

probMAX

24x24 double

	6	7	8	9	10	11	12
1	0	0	0.4700	0.4700	0.0300	0	0
2	0	0	0.0300	0.6300	0.3000	0	0
3	0	0	0	0.1300	0.7700	0.0700	0
4	0	0	0	0.0700	0.4000	0.5000	0
5	0	0	0	0.0300	0.2000	0.5300	0.2000
6	0	0	0	0.0300	0.2000	0.3700	0.2300
7	0	0	0	0.0300	0.0700	0.3000	0.3300
8	0	0	0	0.0300	0.0300	0.2700	0.2700
9	0	0	0.0300	0	0.0700	0.1300	0.2300

Figure 2.3: Part of the sample maxima probability matrix. Each row corresponds to a moving total whereas each column corresponds to a time point (hour) of a day

Using the ranges of the moving totals over multiple days (e.g. 60 days), the mean and standard deviation of the range of each moving total can be computed. The pseudocode of the algorithm for estimating the mean and standard deviation of the range of the hourly energy consumption data are given in Algorithm 2.

Algorithm 2 The pseudocode for estimating the mean and standard deviation of the range of the hourly energy consumption data

1: **Step 1** Obtain the hourly energy consumption of each hour for one day (i.e. the bottom curve in Figure 2.2)

2: **Step 2** Obtain the maximum and minimum

3: **Step 3** Compute the range (maximum minus minimum) and record it

4: **Step 4** Repeat **Step 1** to **Step 3** for each day over multiple days (e.g. 60 days)

5: **Step 5** Compute the mean and standard deviation of the recorded ranges

To compute the mean and standard deviation of the range of the 2-hour moving total, we simply use the *2-hour total energy consumption* at each time point (hour) for one day (i.e. the second to bottom curve in Figure 2.2) in **Step 1** of Algorithm 2. Corresponding modifications to **Step 1** of Algorithm 2 can be made in order to compute the means and standard deviations of the ranges of the 3-hour to 24-hour moving totals.

After applying the above procedure (Algorithm 2 and its modified versions) for all moving totals, the model will then have the means and standard deviations of the ranges of all 24 moving totals. The means and standard deviations of the ranges help represent the regular energy consumption patterns of an occupant and they provide an indication of the normal energy consumptions of the underlying activities.

To sum up, the regular (recurring) energy consumption patterns of an occupant are numerically represented by the two probability matrices (one for the minima and one for the maxima) and the 24 pairs of means and standard deviations.

2.3.2 Final Score and Design Variables

To quantitatively assess the anomalies in the energy pattern of a given day, the energy features of that day need to be provided. As discussed in Section 2.2, the features include the 24 time points of maxima, 24 time points of minima and the 24 ranges.

First, the 24 time points of the maxima (minima) of the given day are converted to scores as described in Algorithm 3. After applying Algorithm 3, the 24 time points of the maxima and 24 time points of the minima will be converted to 48 sub-scores.

Algorithm 3 The pseudocode for converting the 24 times points of the maxima (minima) of a given day to scores

1: **Step 1** Obtain the time of day when the maximum (minimum) occurs for the hourly energy consumption data

2: **Step 2** Retrieve the occurrence probability of that time point being the maximum (minimum) from the maxima (minima) probability matrix

3: **Step 3** Return a positive score (e.g. +1) if the probability is greater than or equal to a pre-defined threshold; otherwise, return a negative score (e.g. -1)

4: **Step 4** Obtain the time of day when the maximum (minimum) occurs for each of the remaining 23 moving totals and repeat **Step 2** and **Step 3**

Second, the 24 ranges of that particular day are converted to scores as described in Algorithm 4.

Algorithm 4 The pseudocode for converting the 24 ranges of a given day to scores

1: **Step 1** Obtain the range of the hourly energy consumption data of the given day

2: **Step 2** Retrieve the mean and standard deviation of the range previously computed as described in Algorithm 2

3: **Step 3** Calculate the Z-Score (standardized value) of the range of the given day

4: **Step 4** Return a positive score (e.g. +1) if the absolute Z-Score is less than or equal to a pre-defined threshold; otherwise, return a negative score (e.g. -1)

5: **Step 5** Obtain the range of each of the remaining 23 moving totals and repeat **Step 2** to **Step 4**

A Z-Score (also know as standard score) measures the distance between an observation and its mean in terms of number of standard deviations [43]. After applying Algorithm 4, the 24 ranges will be converted to 24 sub-scores.

The final score for a given day, which is a quantitative assessment of the anomalies detected in the energy pattern of that day, is the sum of the 72 sub-scores. A more positive final score for a day indicates that the energy pattern on that day has less deviation from the regular energy patterns. Conversely, a more negative final score for a day indicates that there is a large deviation from the regular energy patterns. The pseudocode for the proposed method is provided in Appendix A.

Regardless of the occurrence of anomalies, there are five key design variables in the proposed method that will affect the final score:

1. The selection of the energy consumption data to be included in the model

2. The thresholds for classifying the probabilities
3. The score assignment for the classified probabilities
4. The threshold for classifying the Z-Scores
5. The score assignment for the classified Z-Scores

These five design variables give the model flexibility and they play a significant role in determining the final scores.

2.3.3 Selection of the Electrical Energy Consumption Data

To compute the final score for a given day, some energy consumption data need to be included in the model in order to obtain the regular energy patterns of an occupant. The selection of the energy consumption data is likely to have a significant impact on the probabilities, means and standard deviations that are used to represent the regular (recurring) energy patterns of an occupant. If the regular energy patterns of an occupant did not vary much, the simplest way would be to include as much data in the model as possible. In practice, some changes in the regular energy patterns of an occupant are expected. Including too much data may make it difficult to distinguish between recent regular energy patterns of an occupant and past patterns which may no longer exist. The length of the energy consumption data to be included is one of the design variables that need to be chosen.

2.3.4 Probability Thresholds and Score Assignment

The maxima and minima probability matrices are constructed given the selection of the energy consumption data. The probabilities in the matrices are then classified as probable or not probable using some pre-determined thresholds. The probability is deemed to be probable if it exceeds its corresponding threshold; otherwise, it is deemed to be not probable. The probability matrices are then converted into score matrices. If the actual probabilities are used as the sub-scores for computing the final scores, the sub-scores corresponding to the maxima and minima will have too many different levels; the probable probabilities would not be well differentiated from the not probable probabilities.

Therefore, a probable probability is assigned a positive score (e.g. +1) whereas a not probable probability is assigned a negative score (e.g. -1). Instead of using the actual probabilities, the positive and negative scores assigned are used to compute the final scores as explained in Section 2.3.2. The set of the maxima and minima probability thresholds and the magnitude of the score assignment are two of the design variables that need to be chosen.

2.3.5 Z-Score Threshold and Score Assignment

Using the mean and standard deviation of the range (maximum minus minimum) of each moving total obtained over multiple days (e.g. 60 days), the Z-Score of the range of each moving total on a particular day can be calculated. The Z-Score indicates how far away the range of a moving total on a given day is from its mean in terms of number

of standard deviations. The Z-Scores are then classified as normal or not normal using a pre-determined threshold. The Z-Score is deemed normal if it does not exceed the threshold; otherwise, it is deemed not normal. A normal Z-Score is assigned a positive score (e.g. +1) whereas a not normal Z-Score is assigned a negative score (e.g. -1). The scores assigned are used to compute the final scores as explained in Section 2.3.2. The Z-Score threshold and the magnitude of the score assignment are two of the design variables that need to be chosen.

2.3.6 Flexibility

The flexibility of the model comes from the five design variables mentioned above. The selection of the energy consumption data may affect the deduced regular (recurring) energy consumption patterns of an occupant and hence the final scores. The set of the maxima and minima probability thresholds and the Z-Score threshold determine how frequently an energy pattern has to occur in order to be considered regular. The times of day when maxima and minima of the moving totals occur are more relevant to when activities normally occur, while the Z-Scores of the moving totals are more relevant to the normal energy consumptions of activities. Therefore, the proposed method can place greater emphasis on either the times of day when activities normally occur or the normal energy consumptions of activities by putting more weight on either the probability score assignment or the Z-Score score assignment. An example of the configuration of the design variables is given in Section 3.3.1.

Chapter 3

Validating the Proposed Method with a Survey of Activities

A validation of the effectiveness of the proposed method is presented in this chapter. A survey of activities was conducted to obtain relevant information from the participating lone occupants. The purpose of the survey was to obtain evidence to assess the effectiveness of the method proposed in Chapter 2. The survey of activities, regular activity patterns for the survey, configurations of the design variables and thresholds of acitivty patterns, correlation between energy consumption patterns and activity patterns, comparison of the proposed method and a raw data based approach are discussed below.

3.1 Survey of Activities

To validate the correlation between the household energy consumption data and activity patterns of a person, a survey was conducted to obtain the hourly activities of three lone occupants for a month. The survey was designed to record the activities of daily living (ADLs) of the occupants. The purpose of the survey was to obtain evidence to show that

	00:01 - 01:00	01:01 - 02:00	02:01 - 03:00	03:01 - 04:00	04:01 - 05:00	05:01 - 06:00	06:01 - 07:00	07:01 - 08:00
Not at home								
Sleeping								
Bathing								
Dining at home								
Electric cooking appliances (oven, rice cooker, etc.)								
Dishwasher								
House cleaning (vacuum cleaner, etc.)								
Entertainment (TV, computer, radio, etc.)								
Laundry (washer, dryer, iron, etc.)								
Heater / Air conditioner								
Other energy-consuming activities/appliances* (see note at the bottom of the page)								
	08:01 - 09:00	09:01 - 10:00	10:01 - 11:00	11:01 - 12:00	12:01 - 13:00	13:01 - 14:00	14:01 - 15:00	15:01 - 16:00
Not at home								
Sleeping								

Figure 3.1: A part of the survey timesheet

anomalies in energy consumption patterns would reflect anomalies in activity patterns of the occupants. A part of the survey timesheet is shown in Figure 3.1. Please refer to Appendix B for two complete sample survey timesheets. The listed activities of the survey included the essential ADLs such as sleeping and dining. Each participant was asked to mark the listed activities that took place during any part of each hour on the timesheet.

3.2 Regular Activity Patterns for the Survey

The regular (recurring) activity patterns of each participating occupant were deduced from the activity data of the survey. Using the activity data of the survey, the following can be estimated: 1) the occurrence probability of a certain activity during a day, 2)

the occurrence probability of a certain activity at a given hour, and 3) the occurrence probability of a particular duration of a certain activity during a day. An activity could be considered regular if it is likely to occur during a day. An activity could also be considered regular if it is likely to occur at certain hours or has certain durations during a day. Considering the above conditions, the following features of activity patterns were deduced using appropriate thresholds: 1) highly probable activities during a day, 2) less probable activities during a day, 3) highly probable activities during a given hour, 4) less probable activities during a given hour, and 5) less probable durations of a certain activity during a day. These five features help represent the regular activity patterns of the occupants.

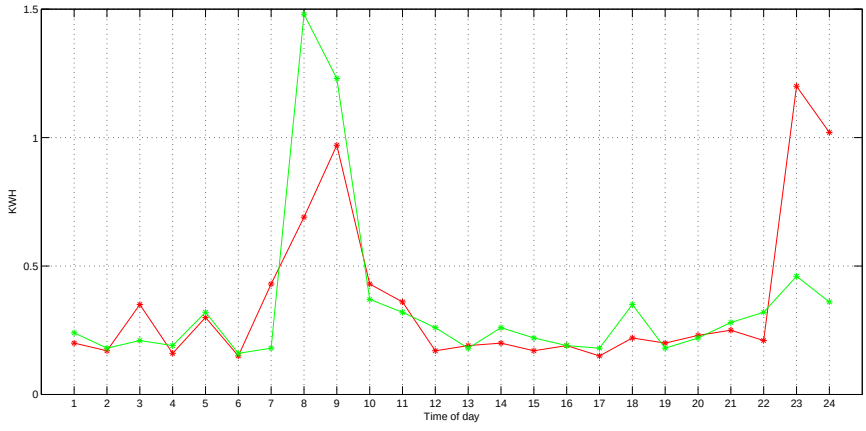

Figure 3.2: Home C - differences in the energy patterns

It was observed that there were occasionally noticeable differences in the energy patterns of two different days although the activities reported in the timesheets of those two days were the same or similar. For instance, the hourly energy patterns of Home C on

March 2 (red) and March 15 (green) are shown in Figure 3.2 (please refer to Appendix B for the complete timesheets of the activities). Although the activities from 7AM to 9AM on those two days were the same, there were noticeable differences in the two energy patterns during the same time period. This indicates that the activity patterns might not only involve the times of day when activities occur but also the energy consumption related to the activities. However, the energy consumption data of the individual home electric appliances were not explicitly available. Therefore, the daily energy consumptions (i.e. the total energy consumption of all appliances during a day) were chosen to help represent the regular activity patterns of the occupants.

Therefore, the six features that were used to quantitatively assess anomalies in activity patterns of the occupants are:

1. The daily highly probable activities
2. The daily less probable activities
3. The hourly highly probable activities
4. The hourly less probable activities
5. The daily less probable durations of activities
6. The less probable daily energy consumption

The quantitative assessment of anomalies in a daily activity pattern is called *level of suspicion* in this thesis. The above six features and the level of suspicion are described below.

3.2.1 Daily Highly Probable Activities

An activity is classified as daily highly probable if the occurrence probability of the activity during a day is greater than or equal to a given threshold. A non-occurrence of a daily highly probable activity during a given day will increase the level of suspicion of that day. The magnitude of the increase is equal to the expected duration of the activity during a day rounded to the closest integer. For instance, if the expected duration of an activity is eight hours and it does not occur on a given day, the level of suspicion of that day will be increased by eight.

3.2.2 Daily Less Probable Activities

An activity is classified as daily less probable if the occurrence probability of the activity during a day is less than or equal to a given threshold. An occurrence of a daily less probable activity during a given day will increase the level of suspicion of that day. The magnitude of the increase is equal to the expected duration of the activity during a day rounded to the closest integer.

3.2.3 Hourly Highly Probable Activities

For each activity, an hour is classified as highly probable if the occurrence probability of the activity during that hour is greater than or equal to a given threshold. That an activity occurs during a given day but does not occur at the highly probable hours will increase the level of suspicion of that day. The magnitude of the increase is equal to the

number of missed highly probable hours. For instance, if the highly probable hours of a given activity are 8AM and 9AM but the activity occurs only at 8AM on a particular day, the level of suspicion of that day will be increased by one. The level of suspicion of a given day will not be increased due to missed highly probable hours if the activity does not occur at all on that day; the hourly highly probable feature is relevant only if the activity occurs during a day. If the activity is deemed to be daily highly probable, the level of suspicion will have been increased due to its absence during the given day.

3.2.4 Hourly Less Probable Activities

For each activity, an hour is classified as less probable if the occurrence probability of the activity during that hour is less than or equal to a given threshold. That an activity occurs at the less probable hours will increase the level of suspicion of the day. The magnitude of the increase is equal to the number of occurred less probable hours. For example, if the less probable hours of a given activity are 3AM and 4AM and the activity occurs at 3AM on a given day, the level of suspicion of that day will be increased by one.

3.2.5 Daily Less Probable Durations of Activities

For each activity, a duration of the activity during a day is classified as less probable if the occurrence probability of that duration is less than or equal to a given threshold. An occurrence of a less probable duration of an activity will increase the level of suspicion of the day. The magnitude of the increase is equal to the absolute difference (in hours)

between the actual duration and the most probable duration of the activity. The level of suspicion of a day will not be increased due to a daily less probable duration if the activity does not occur at all on that day; this feature is relevant only if the activity occurs during a day. If the activity is deemed to be daily highly probable, the level of suspicion will have been increased due to its absence during the given day.

3.2.6 Less Probable Daily Energy Consumption

Given the daily energy consumption data for multiple days (e.g. 30 days), the mean daily energy consumption and the standard deviation can be calculated. The Z-Score of the daily energy consumption can then be computed. A Z-Score measures the distance between an observation and its mean in terms of number of standard deviations. The Z-Score of the daily energy consumption is classified as less probable if its absolute value is greater than or equal to a given threshold. An occurrence of a less probable daily Z-Score will increase the level of suspicion of the day. The less probable daily energy consumption is deemed to be a day-long anomaly, and the magnitude of the increase is therefore 24.

3.2.7 Level of Suspicion

The accumulated level of suspicion value of a day is equal to the sum of the level of suspicion values from the above six different features. The level of suspicion is meant to be a quantitative assessment of anomalies in a daily activity pattern of an occupant. A

high level of suspicion value of a day indicates a large deviation from the regular activity patterns.

3.3 Configurations of the Design Variables and Thresholds of Acitivty Patterns

The design variables for the proposed method (i.e. selection of the electrical energy consumption data, probability thresholds and score assignment and z-score threshold and score assignment) need to be determined before a score can be computed. The thresholds of the above-mentioned six features of activity patterns for the survey also need to be set before a level of suspicion can be calculated. The configurations of the design variables for the proposed method and thresholds of regular activity patterns for the survey are discussed in this section.

3.3.1 Design Variables for the Proposed Method

The design variables for the proposed method affect how often an energy pattern needs to occur in order to be considered regular (recurring), and they play significant roles in assessing anomalies in energy patterns. The scheme for choosing the design variable values is discussed in Appendix C. The configurations of the design variables are shown in Table 3.1. Due to the limited amount of energy data, all scores of Home A were computed based on the same 30 days of energy consumption data. In other words, the

30-day data window was the same (static) for all scores.

For Home B and Home C, the score of a given day was computed based on 60 days of energy consumption data immediately before. In other words, the 60-day data window was different (dynamic) for each score. Using the given configurations, the maximum possible score is 72 while the minimum possible score is -72.

	Home A	Home B	Home C
Length of Data (days)	30 (static)	60 (dynamic)	60 (dynamic)
Probability Threshold			
-Occurrence of Max	0.1	0.1	0.15
-Occurrence of Min	0.15	0.1	0.15
Score Assignment	+/-1	+/-1	+/-1
Range Z-Score Threshold	1.5	1	1
Score Assignment	+/-1	+/-1	+/-1

Table 3.1: The configurations of the design variables for the proposed method

3.3.2 Thresholds of Regular Activity Patterns for the Survey

The thresholds of regular activity patterns affect the regular activity patterns and consequently the assessments of anomalies in activity patterns. The threshold values used in this study are shown in Table 3.2. The length of survey data which the various probabilities were based on and the length of daily energy consumption data which the daily Z-Scores were based on are shown in Table 3.3. Due to the limited amount of data, the

daily Z-Scores for Home A were computed based on the available 30 days of daily energy consumption data. The daily Z-Scores for Home B and Home C were computed based on one year of daily energy consumption data. The scheme for choosing the threshold values is discussed in Appendix D.

Features of Activity Patterns	Home A	Home B	Home C
	Probability		
Daily highly probable activities	$\geq$0.8	$\geq$0.8	$\geq$0.9
Daily less probable activities	$\leq$0.2	$\leq$0.2	$\leq$0.2
Hourly highly probable activities	$\geq$0.9	$\geq$0.8	$\geq$0.8
Hourly less probable activities	$\leq$0.05	$\leq$0.05	$\leq$0.05
Daily less probable durations of activities	$\leq$0.05	$\leq$0.05	$\leq$0.05
	|Z-Score|		
Less probable daily energy consumptions	$\geq$1	$\geq$0.95	$\geq$1.1

Table 3.2: The threshold values of regular activity patterns for the survey

	Home A	Home B	Home C
Length of	Days		
Survey data	30	30	31
Daily energy consumption data	30	365	365

Table 3.3: The lengths of the survey and daily energy consumption data in days

3.4 Correlation between Energy Consumption Patterns and Activity Patterns

In Chapter 2, the *Score*, a quantitative assessment of anomalies in an energy consumption pattern, was introduced. In this chapter, the *Level of Suspicion*, a quantitative assessment of anomalies in an activity pattern, was introduced. In this section, we show the correlation between the Score (energy consumption patterns) and the Level of Suspicion (activity patterns) for each participating occupant. The scores sorted in descending order of the level of suspicion values of Home A, Home B and Home C are shown in Figure 3.3, Figure 3.4 and Figure 3.5 respectively. In Figure 3.3, the level of suspicion values of the first nine scores range from 57 down to 24 while the level of suspicion values of the remaining scores range from 12 down to 0. In Figure 3.4, the level of suspicion values of the first two scores are 48 and 26 while the level of suspicion values of the remaining scores range from 11 down to 0. In Figure 3.5, the level of suspicion values of the first eigth scores range from 30 down to 24 while the level of suspicion values of the remaining scores range from 7 down to 0. Although the score does not monotonically increase as the level of suspicion value decreases, the scores tend to be higher when the level of suspicion values are lower. In other words, the score seems somewhat correlated with the level of suspicion.

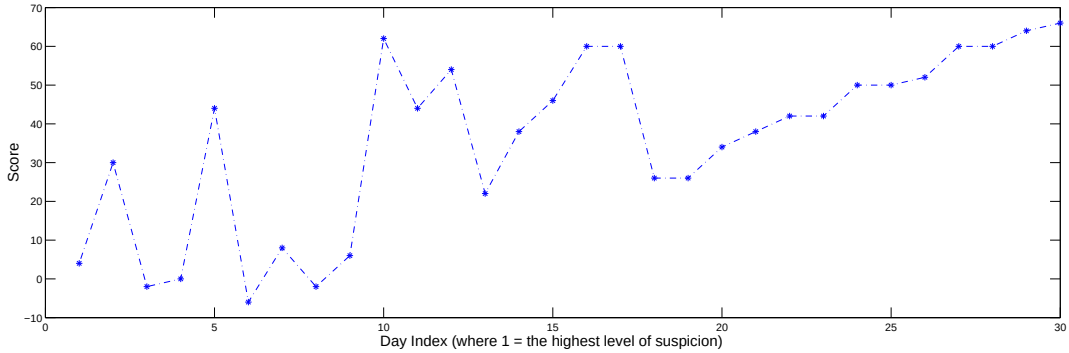

Figure 3.3: Home A - scores sorted in descending order of level of suspicion values

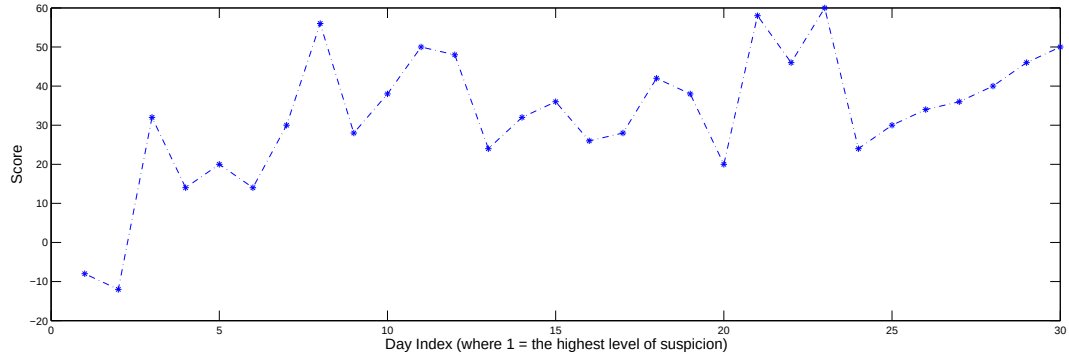

Figure 3.4: Home B - scores sorted in descending order of level of suspicion values

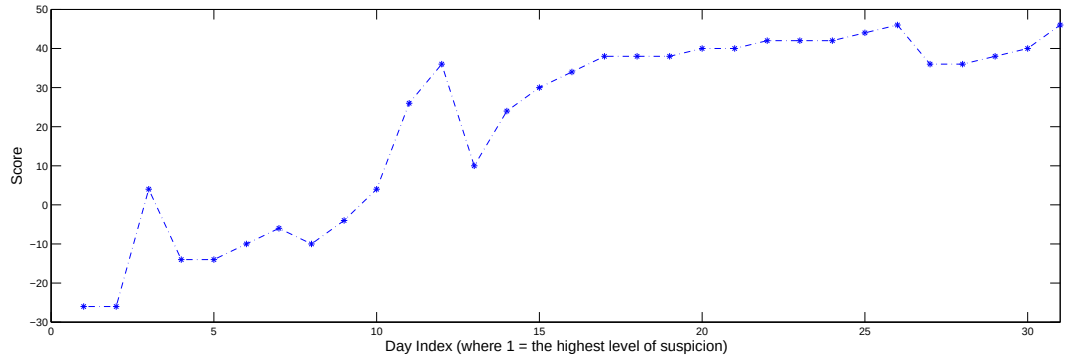

Figure 3.5: Home C - scores sorted in descending order of level of suspicion values

3.5 Comparison of the Proposed Method and a Raw Data Based Approach

The objective of this research work is to classify the daily activity patterns of an occupant as regular or irregular using the corresponding energy consumption patterns. A simple way would be to measure the similarities of the raw energy consumption sequences to one another and group similar energy sequences into distinct clusters (i.e. to do a clustering of the raw energy consumption sequences). Each cluster can then be assigned a class label (i.e. regular or irregular). This can be considered a raw data based clustering approach. However, the raw energy consumption data without any further processing may not be an effective indicator of the underlying activity patterns.

The proposed method infers anomalies in activity patterns using the features extracted from the raw energy consumption data and the model built using the features. The proposed method thus uses a mixture of feature based and model based approach. This section shows that the proposed method is more accurate than a chosen raw data based approach for classifying activity patterns of the occupants. The level of suspicion, which was deduced from the activity survey, is used as the pseudo ground truth for the comparison.

Instead of attempting to deduce whether an activity pattern is normal or not normal from the activity data, a simpler way would be to directly ask the occupant. The answer from the occupant could then be used as the ground truth. However, the definition of

"normal" can vary from person to person. Therefore, we would want to ask the occupant a series of questions so that the answers might tell us whether the activity pattern of a given day is normal or not normal according to our chosen definition.

The survey of this research work attempted to ask the occupant to rate his/her mood and health respectively on a 5-point scale. The change in the ratings from one day to another might indicate the change in his/her wellness (probably the activity patterns as well). Unfortunately, the participating occupants picked the same rating every day, which did not provide much useful information to this research work. Therefore, we decided to deduce the pseudo ground truth (based on the level of suspicion) from the activity data of the survey instead.

The pseudo ground truth from the survey, training sets and test sets, clustering of the scores provided by the proposed method, clustering of the raw energy consumption sequences, and performance evaluation are provided in this section.

3.5.1 Pseudo Ground Truth from the Survey

The pseudo ground truth of whether the daily activity patterns of the participating lone occupants were regular or irregular was deduced from the level of suspicion values. Although a higher level of suspicion value indicates that the daily activity pattern of an occupant is more irregular, the boundary between the higher level of suspicion values and the lower level of suspicion values of each participating occupant has not been defined. Fuzzy C-Means (FCM) clustering [44] was used to cluster the level of suspicion values

into two groups: regular (low level of suspicion) and irregular (high level of suspicion). FCM was chosen because it is well-known and works well with low dimensional data. A day was then labelled as regular or irregular according to which resulting cluster it belonged to. The cluster with a lower center value (defined as the mean of all data points that belongs to the cluster) was considered the regular cluster while the cluster with a higher center value was considered the irregular cluster.

3.5.2 Training Sets and Test Sets

For both the scores and raw energy sequences, the available data of each home were split into a training set and a test set. There were only 30 days (31 days for Home C) that have the pseudo ground truth data because the survey was only conducted for a month. In order to compare the clustering results of the proposed method and the chosen raw data based approach based on the pseudo ground truth, we were limited to compare only 30/31 days of the results deduced from the energy consumption patterns by the proposed method and the raw data based approach. Therefore, for each home, 20 consecutive days (21 days for Home C) of data were chosen as the training data (i.e. either the first 20/21 days or the last 20/21 days of data) while the remaining 10 consecutive days were chosen as the test set data. The lengths (number of samples) of the training set and test set are shown in Table 3.4.

	Training Set (days)	Test Set (days)
Home A	20	10
Home B	20	10
Home C	21	10

Table 3.4: The lengths of the training set and test set data

3.5.3 Clustering of the Scores Provided by the Proposed Method

Although a lower score indicates that the energy consumption pattern of an occupant on a given day is more irregular, the boundary between the lower scores and higher scores for each participating occupant needs to be defined. First, FCM was used to cluster the data of each training set into two groups respectively. The cluster with a lower center value was considered the irregular cluster while the cluster with a higher center value was considered the regular cluster. Second, each data point of each test set was classified as regular or irregular depending on which cluster's center it was closer to.

3.5.4 Clustering of the Raw Energy Consumption Sequences

The proposed method requires an extra step to extract features from the raw energy consumption data and build a model using the features before the clustering. The additional step would be questionable if the proposed method does not perform better than a raw data based approach. In [32], it was found that the Euclidean distance measure is the

overall best similarity measure for clustering hourly energy consumption data among the other three well-known measures (Mahalanobis distance [33], Dynamic Time Warping distance [34] and Pearson's correlation [35]). Since FCM was the clustering technique used in [32], FCM with Euclidean distance measure was chosen for clustering the raw energy consumption sequences in this thesis.

First, FCM was used to cluster the energy consumption sequences of each training set six times, each time with a different number of pre-determined clusters (2 to 7). Second, the Dunn index [36] was used to determine the best number of clusters for each training set according to the six different clustering results. The Dunn index is basically the ratio of the minimum distance between clusters to the maximum distance between data points within the same cluster and should be maximized [45]. According to the Dunn index, the best number of clusters was 3 for all Home A, Home B and Home C. Therefore, FCM was used again to cluster the energy consumption sequences of each training set for each home into three clusters respectively. Third, the three clusters for each home were labelled as regular or irregular respectively. Two clusters were to be assigned the same class label. The particular combination of the three labelled clusters which yielded the best result when comparing against the pseudo ground truth was chosen. Last, each energy consumption sequence of each test set for each home was classified as regular or irregular depending on which cluster's center it was closer to. The distance measure used for the classification of the test set data was again Euclidean distance measure.

3.5.5 Performance Evaluation

The clustering results of the scores and the raw energy sequences are compared here. The pseudo ground truth was derived from the clustering result of the level of suspicion values as described in Section 3.5.1. Flowcharts of the method for deriving the pseudo ground truth from an activity perspective, the proposed method based on FCM and the raw data based approach are shown in Figure 3.6.

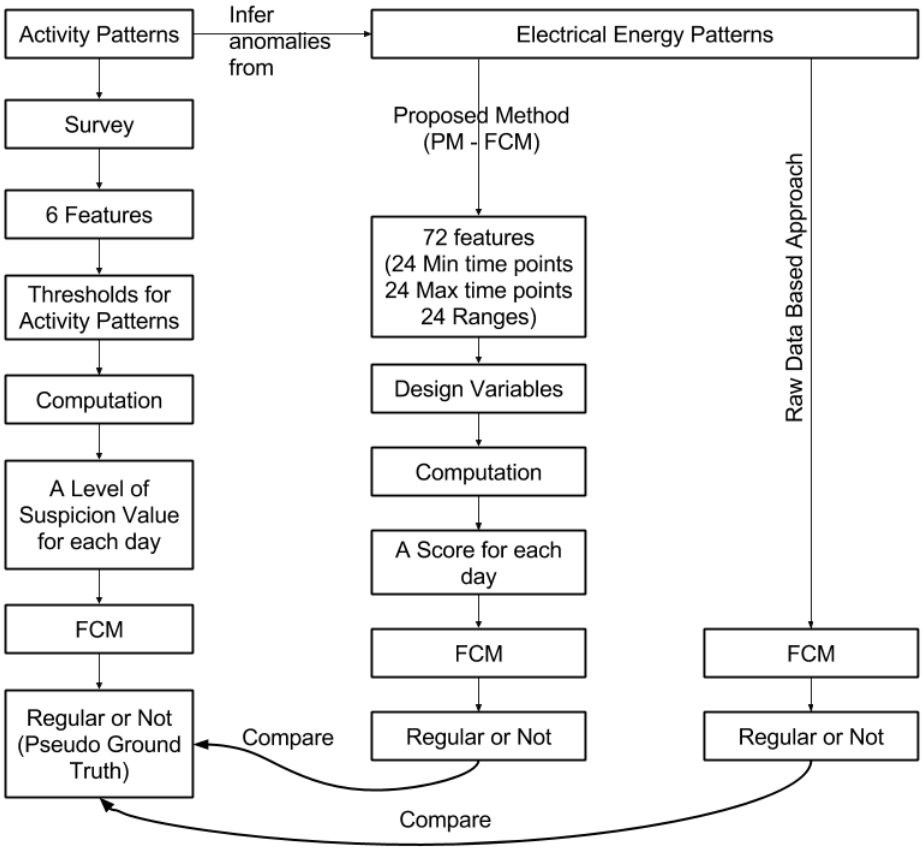

Figure 3.6: Flowcharts of the method for classifying a day as regular or irregular from an activity perspective (left), the proposed method based on FCM (center), and the raw data based approach (right)

The performances of the proposed method and the chosen raw data based approach are summarized in Table 3.5. The performance evaluation was based on all three datasets of the three lone occupants. The performance shown here is the average performance for all Homes A, B and C. The number of missed detections indicates the number of irregular days that were incorrectly labelled as regular by the proposed method or the raw data based approach. The number of false alarms indicates the number of regular days that were incorrectly labelled as irregular by the proposed method or the raw data based approach. Sensitivity is the proportion of irregular days that were correctly labelled by the proposed method or the raw data based approach. Specificity is the proportion of regular days that were correctly labelled by the proposed method or the raw data based approach.

	Proposed Method	Raw Data Based Approach [32]
Training Set		
Accuracy	57/61 (93.44%)	54/61 (88.52%)
Missed Detection	1	1
False Alarm	3	6
Sensitivity	13/14 (92.86%)	13/14 (92.86%)
Specificity	44/47 (93.62%)	41/47 (87.23%)
Test Set		
Accuracy	26/30 (86.67%)	22/30 (73.33%)
Missed Detection	1	3
False Alarm	3	5
Sensitivity	4/5 (80%)	2/5 (40%)
Specificity	22/25 (88%)	20/25 (80%)

Table 3.5: The performances of the proposed method and the chosen raw data based approach

As described in Section 3.2.1 to Section 3.2.7, the level of suspicion was deduced from the five features derived from the survey of activities and the daily energy consumption data (i.e. the total energy consumption of all appliances during a day). Unlike the other five features, the daily energy consumption data were not derived from the survey of activities. The reason why daily energy consumptions were included is discussed in Section 3.2. To examine the significance of the daily energy consumptions in the performance

evaluation, a test was conducted to evaluate the performances of the proposed method and the raw data based approach when the pseudo ground truth did not consider the daily energy consumptions (please refer to Appendix E for details). Removing the less probable daily energy consumption feature from the pseudo ground truth appeared to adversely affect the accuracy performances of both the proposed method and the raw data based approach equally. The training set accuracies of the proposed method and the raw data based approach dropped by 19.30% and 20.37% respectively, while the test set accuracies dropped by 15.38% and 13.64% respectively.

3.6 Discussion

The numerical results show that anomalies in energy consumption patterns detected by the proposed method tend to correlate with anomalies in activity patterns. The results also show that the proposed feature and model based method outperforms the raw data based approach. It indicates that the features extracted from the energy consumption data better represent anomalies in the underlying activities than the raw energy consumption data. Although the training set and test set available are relatively small, the performance of the proposed method is quite encouraging. Given the sensitivity and specificity, the proposed method can detect the anomalies in activity patterns of the occupants quite well. The proposed method does not identify whether the anomalies detected indicates a positive or negative change in the occupant's wellness. However, it can trigger an alert to caregivers or relatives who can then investigate in a timely manner.

Chapter 4

Reducing the Energy Features Based on mRMR Feature Selection

As mentioned in Section 2.2, the energy consumption pattern of a given day is represented by 24 time points of maxima, 24 time points of minima, and 24 ranges. Thus, there is a total of 72 different features for representing the energy consumption pattern of a day. It has been shown that using subsets of features might improve the performances of classifiers for various classification tasks [46–48]. In this chapter, we show the performance of the proposed method when subsets of the 72 energy features were considered. An overview of feature selection methods, the minimum Redundancy Maximum Relevance (mRMR) feature selection, and the performance evaluation are provided in the following sections.

4.1 Overview of Feature Selection Methods

Feature selection methods are typically used for the following reasons: 1) simplifying models for providing better interpretations of the underlying processes that generated the

data, 2) reducing training times of classifiers, 3) improving classification accuracy [49]. Feature selection methods generally help improve classification accuracy by: 1) leaving out irrelevant features (noise) and 2) reducing overfitting. Feature selection methods can generally be divided into three main categories: 1) filters, 2) wrappers and 3) embedded methods.

Filter type methods rank features or select subsets of features based on some chosen conditions (e.g. mutual information, Pearson's correlation [35]) and are independent to classifiers. Thus, filter type methods filter out irrelevant features before the learning process occurs. Some examples of filter methods are RELIEF [50] and mRMR [39, 40]. Filters are generally faster than wrappers and embedded methods because filters rank features according to their characteristics and are independent of the classification algorithms.

Wrapper type methods select a subset of features and use a classifier (e.g. Naive-Bayes classifier, Support Vector Machine [51]) to assess the quality of the subset. It is generally impractical to exhaustively search through every combination of all considered features. Two typical strategies are forward selection and backward elimination [52]. Forward selection begins a search with an empty set and adds one feature at a time while backward elimination begins a search with a full set of all considered features and eliminates one feature at a time. Wrappers generally yield better classification performance than filters at the expense of high computational cost.

Embedded methods are similar to wrapper type methods except that feature selec-

tions are part of the classifiers and they cannot be separated. In contrast, wrapper type methods can be combined with any classifier. Some examples of embedded methods are Support Vector Machine Recursive Feature Elimination (SVM-RFE) [53] and Locally Weighted Naive-Bayes classifier [54]. Embedded methods are generally less computationally intensive than wrappers.

A filter type method was adopted in this thesis because there were 72 energy features and hence the computational cost was a concern. Most filter type methods select features based on their relevance to target classes but do not consider the redundancy among features [39, 40]. The minimum Redundancy Maximum Relevance (mRMR) feature selection [39, 40], which considers both relevance of features to target classes and redundancy among selected features, was therefore adopted.

4.2 mRMR Feature Selection

The mRMR feature selection [39, 40] is a filter type method which ranks features according to mutual informations between target classes (i.e. regular and irregular for our classification problem) and individual features (relevance) and mutual informations between selected features (redundancy). The idea of mRMR is to select a subset of features which are maximally dissimilar (minimum redundancy) while maximizing the mutual information between selected features and target classes (maximum relevance).

Let S denote the set of selected features. The minimum redundancy condition [40] is

$$\min W_I, \quad W_I = \frac{1}{|S|^2} \sum_{i,j \in S} I(i,j), \tag{4.1}$$

where $I(i,j)$ represents the mutual information between feature i and feature j, and $|S|$ is the number of selected features in S. Let h denote the class variable (i.e. regular or irregular for our classification problem). The maximum relevance condition [40] is

$$\max V_I, \quad V_I = \frac{1}{|S|} \sum_{i \in S} I(h,i), \tag{4.2}$$

where $I(h,i)$ represents the mutual information between feature i and class h, and $h = \{h_1, h_2, ..., h_K\}$ if there are K different classes.

The objective of mRMR is to optimize the above two conditions simultaneously. It requires a criterion for combining the two conditions. Therefore, two criteria have been considered [40]: the Mutual Information Difference (MID) and Mutual Information Quotient (MIQ) criteria.

$$MID, \quad \max(V_I - W_I), \tag{4.3}$$

$$MIQ, \quad \max(V_I / W_I), \tag{4.4}$$

In mRMR, the feature with the highest $I(h,i)$ is chosen as the first feature. The remaining features are chosen incrementally. The first selected feature is ranked first, the second selected feature is ranked second, and so forth. Suppose we have included $m-1$ features in S_{m-1} (where $m > 1$) and want to add the mth feature. Let S'_{m-1} be the set of features that have not been selected. The feature j in S'_{m-1} which maximizes the MID

or MIQ condition is selected as the mth feature [39]:

$$MID, \quad \max_{j \in S'_{m-1}} \left[I(j,h) \; - \; \frac{1}{m-1} \sum_{i \in S_{m-1}} I(j,i) \right], \tag{4.5}$$

$$MIQ, \quad \max_{j \in S'_{m-1}} \left[I(j,h) \; / \; [\frac{1}{m-1} \sum_{i \in S_{m-1}} I(j,i)] \right]. \tag{4.6}$$

One needs to determine which one of the two criteria to use when using mRMR. After several test runs, it was found that using either MID or MIQ criterion eventually yielded similar classification accuracy for the proposed method. As neither MID nor MIQ had apparent advantage over each other, the MIQ criterion was adopted for the performance evaluation.

4.3 Performance Evaluation

To use mRMR, we needed to provide the class label of each sample (day) and the 72 energy features of each sample. First, each sample was labelled as 0 (regular) or 1 (irregular) according to the pseudo ground truth introduced in Section 3.5.1. Second, the 72 sub-scores which correspond to the 72 energy features (as mentioned in Section 2.3.2) of each sample were retrieved. Third, mRMR was used to rank the energy features for Homes A, B and C respectively. Please refer to Appendix F for the detailed rankings. The top ranked feature is the feature i that maximizes the relevance condition (i.e. Equation 4.2). The next highest ranked feature is the feature j that maximizes the MIQ condition (i.e. Equation 4.6). A flowchart of the proposed method based on mRMR and FCM

(PM-mRMRnFCM) is shown in Figure 4.1. A flowchart of the proposed method based on FCM (PM-FCM) is also shown in Figure 4.1 for comparison. PM-mRMRnFCM is essentially the same as PM-FCM, except that PM-mRMRnFCM reduces the number of energy features before the FCM clustering.

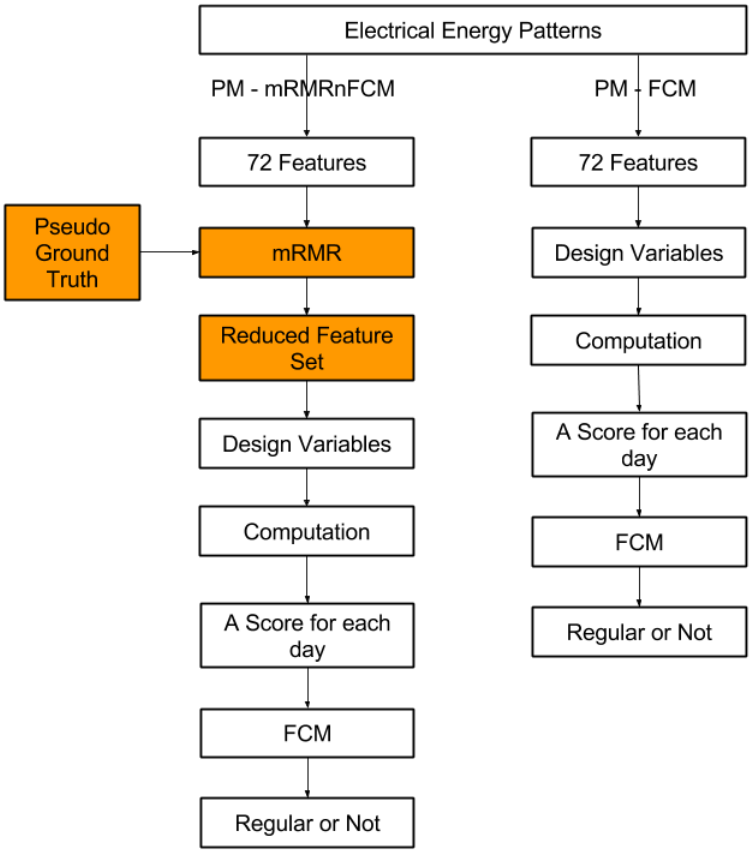

Figure 4.1: Flowcharts of the proposed method based on mRMR and FCM (left) and the proposed method based on FCM (right)

The performance of this proposed method was then evaluated when a subset of the top ranked m (where $1 \leq m \leq 72$) feature(s) was considered. The performance evaluation

was based on all three datasets of Homes A, B and C. Please note that each home has its own ranking of the 72 energy features. The top ranked m feature(s) means the top ranked m feature(s) of each home respectively. The performance shown here is the average performance for all Homes A, B and C. As can be seen in Figure 4.2, neither the best training set accuracy nor the best test set accuracy was achieved when all 72 features were considered. The training set accuracy was the highest when the top ranked four to 12 features were considered. The test set accuracy was the highest when the top ranked two or 13 to 60 features were considered. The accuracy percentages of some key points in Figure 4.2 are listed in Table 4.1.

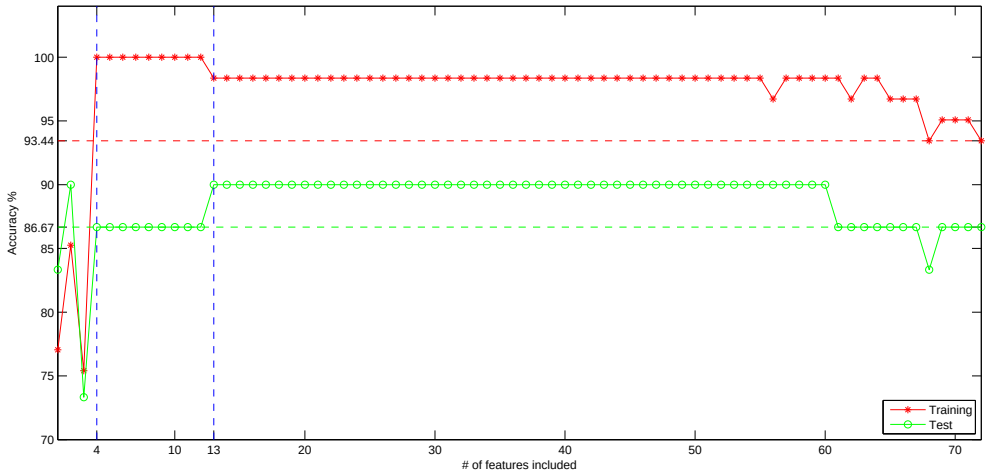

Figure 4.2: Accuracies of the proposed method when subsets of the respective top ranked one to 72 feature(s) of Homes A, B and C were considered

# of feature(s)	Training accuracy	Test accuracy
1	77.05%	83.33%
4	100%	86.67%
13	98.36%	90%
72	93.44%	86.67%

Table 4.1: The performances of the proposed method when subsets of the top ranked one, four, 13 and 72 feature(s) were considered

4.4 Discussion

It was found that the performance of the proposed method could be improved when subsets of the 72 energy features were considered based on mRMR feature selection. When the top ranked four features were considered, the training set accuracy was improved as compared to when all features were considered. When the top ranked 13 features were considered, both the training set and test set accuracies were improved as compared to when all features were considered.

As can be seen in Table 4.1, the training accuracy was lower than the test accuracy when only the top ranked feature was considered. First, the learning algorithm might not be able to capture the underlying trend of the data if too few features are included (under-fitting). Second, there are more training samples than test samples. Therefore, a under-fitted model might occasionally fit the test samples better than the training samples.

As can be seen in Figure 4.2, the training and test accuracies at both ends of the figure do not seem very stable. It is because mRMR, a filter type feature selection method, ignores the effect of the selected subset of features on the classification accuracy for ranking features. Therefore, there is no guarantee that the classification accuracy will be improved when the next highest ranked feature is included.

In [55, 56], the similar effects on the classification accuracy of some filter type feature selection methods (e.g. information gain, mRMR) can also be observed. In [55], the researchers classified some Comparative Genomic Hybridization (CGH) data based on mRMR and SVM. The instability in the classification accuracy can be observed when the number of selected features was changed. The instability might be more noticeable for our classification problem because we have a relatively small sample size; each sample in the training set accounts for 1.64% ($\frac{1}{61} \times 100\%$) of the classification accuracy and each sample in the test set accounts for 3.33% ($\frac{1}{30} \times 100\%$) of the classification accuracy.

In [56], the researchers investigated the relationship between feature selection (and extraction) methods and the resulting classification accuracy based on various classifiers such as SVM. From the experimental results in [56], the following can be observed:

1. The classification accuracy was not stable when the number of selected features (ranked by a given method) was changed.
2. The effectiveness of a subset of selected features (ranked by a given method) on the classification accuracy might vary from classifier to classifier.
3. There are generally no rules for choosing the optimal number of features.

Chapter 5

Classifying Electrical Energy Consumption Patterns Based on C-ELM

In Chapter 2 and Chapter 3, the Proposed Method based on Fuzzy C-Means clustering (PM-FCM) was introduced. PM-FCM classifies energy patterns without labelled training data and is hence an unsupervised learning technique. In Chapter 4, PM-FCM was enhanced by adopting the mRMR feature selection (PM-mRMRnFCM). The mRMR feature selection ranks the 72 energy features according to some labelled data based on the pseudo ground truth deduced from the survey (introduced in Section 3.5.1). PM-mRMRnFCM classifies energy patterns using subsets of energy features according to the ranking provided by mRMR. The training process of PM-mRMRnFCM is still unsupervised because the training process does not require labelled training data. In this chapter, we propose a supervised learning technique based on the Curious Extreme Learning Machine (C-ELM) [41] which uses both the energy features and the pseudo ground truth (i.e. labelled training data) for training. The classification based on C-ELM and its

performance evaluation are discussed below.

5.1 Classification Based on C-ELM

In this section, a supervised learning technique is adopted for classifying energy consumption patterns. The Curious Extreme Learning Machine (C-ELM) [41] is chosen because it has been shown that C-ELM outperforms other popular classifiers such as SVM based on some benchmark classification problems [41]. C-ELM requires the input vectors to be labelled by the desired outputs for training. For our classification problem, the input vectors are the 72-dimensional vectors of the energy features, each labelled by the desired output (i.e. regular or irregular) which is the pseudo ground truth deduced from the survey. The pseudocode for the C-ELM training is given in Algorithm 5. The details of C-ELM are given in [41].

Algorithm 5 The pseudocode for the C-ELM training

1: **Step 1** Present a sample $(\mathbf{x}^t, c^t)$, where $\mathbf{x}^t$ denotes the t-th input vector (i.e. the 72 energy features) and c^t denotes its class label (i.e. regular or irregular)

2: **Step 2** Compute the values of four variables (*Novelty*, *Uncertainty*, *Conflict*, and *Surprise*) for the given input data

3: **Step 3** Based on the four variables in **Step 2**, select one of the following learning strategies: 1) add a hidden neuron, 2) delete a hidden neuron, and 3) update network parameter

4: **Step 4** Increment t by 1

5: **Step 5** Repeat **Step 1** to **Step 4** until the last sample has been reached

C-ELM also requires choosing the following five thresholds for the four variables:

1. The low threshold for *Novelty*
2. The high threshold for *Novelty*
3. The threshold for *Uncertainty*
4. The threshold for *Conflict*
5. The threshold for *Surprise*

Each threshold value ranges from 0 to 1 and hence it is not practical to exhaustively explore all possible combinations. However, about 10000 different combinations of the five threshold values were tested. The chosen threshold values are those that provided the best testing accuracy by experiment. The threshold values for Homes A, B and C are shown in Table 5.1.

Threshold	Home A	Home B	Home C
Novelty Low	0.1	0.1	0.1
Novelty High	0.3	0.5	0.4
Uncertainty	0.2	0.2	0.3
Conflict	0.1	0.2	0.5
Surprise	0.6	0.3	0.6

Table 5.1: The threshold values for C-ELM for Homes A, B and C

The lengths (number of samples) of the training set and test set for each home are the same as shown in Table 3.4 in Section 3.5.2.

5.2 Performance Evaluation

The performance of the Proposed Method based on C-ELM (PM-CELM) was evaluated. PM-CELM was then enhanced by adopting mRMR (PM-mRMRnCELM). Flowcharts of PM-CELM and PM-mRMRnCELM are shown in Figure 5.1.

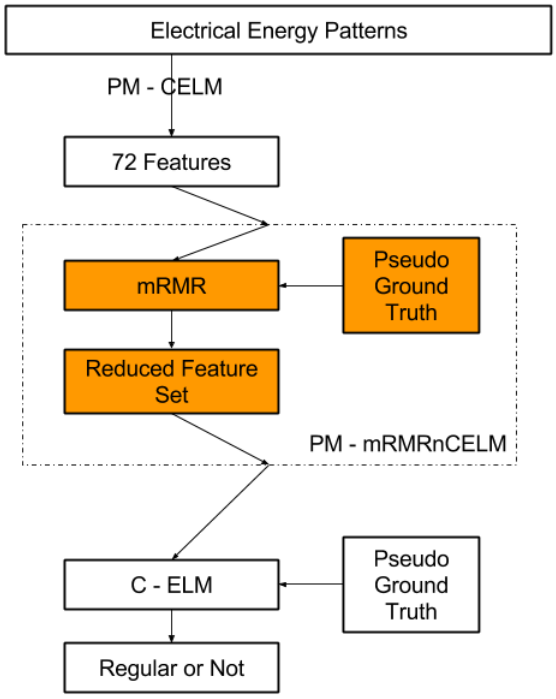

Figure 5.1: Flowcharts of the proposed method based on C-ELM (without the shaded steps) and the proposed method based on mRMR and C-ELM (with the shaded steps)

The accuracy performance of PM-CELM is shown in Table 5.2. The performances of PM-FCM (as in Table 3.5) and PM-mRMRnFCM (as in Table 4.1) are also shown here for comparison. The performance shown here is the average performance for all Homes A, B and C.

	PM-CELM	PM-FCM	PM-mRMRnFCM	
			4 features	13 features
Training Set Accuracy	100%	93.44%	100%	98.36%
Test Set Accuracy	96.67%	86.67%	86.67%	90%

Table 5.2: The performances of PM-CELM, PM-FCM and PM-mRMRnFCM

The procedure for adopting mRMR is the same as discussed in Section 4.3. Since mRMR is independent of the learning algorithm used, the rankings of the energy features (please refer to Appendix F) are exactly the same for both PM-mRMRnCELM and PM-mRMRnFCM. The performance of PM-mRMRnCELM was evaluated and is shown in Figure 5.2. As can be seen in Figure 5.2, the best training accuracy and test accuracy are 100% and 96.67% respectively. The accuracy percentages of some key points in Figure 5.2 are listed in Table 5.3. The best performance was achieved when all 72 features were considered. However, the same performance could also be achieved when only the top ranked 11 features were considered. As can be seen in Figure 5.2, the classification accuracy does not seem very stable. The instability in the classification accuracy was discussed in Section 4.4.

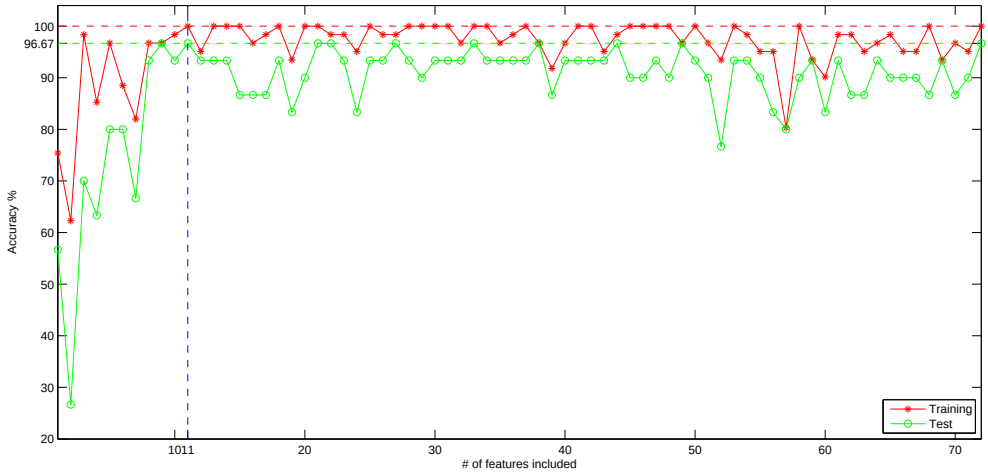

Figure 5.2: Accuracies of PM-mRMRnCELM when subsets of the top ranked one to 72 feature(s) of Homes A, B and C were considered

# of features	Training accuracy	Test accuracy
11	100%	96.67%
72	100%	96.67%

Table 5.3: The performances of PM-mRMRnCELM when subsets of the top ranked 11 and 72 features were considered

5.3 Discussion

In this chapter, we proposed a supervised learning method based on C-ELM (PM-CELM). It was then enhanced by adopting the mRMR feature selection (PM-mRMRnCELM). They provide two simple solutions for classifying energy consumption patterns as either regular or irregular if the ground truth (of whether the underlying activity pattern is

regular or irregular) is available for training. On the other hand, PM-FCM, the proposed unsupervised learning method based on FCM, as described in Chapter 2 and Chapter 3, can operate without the ground truth or labelled training data. The performance evaluation in this chapter shows that PM-CELM outperforms PM-FCM, as might be expected. However, PM-CELM can only be used when labelled training data are available.

Chapter 6

Conclusion and Future Work

In this chapter, the main findings are summarized and some topics of future research are discussed.

6.1 Conclusion

This thesis explored the correlation of household electricity consumption patterns and activity patterns of an occupant. The objective of the research work in this thesis is to classify the daily activity patterns of an occupant as regular or irregular. The fundamental assumption was that anomalies in energy consumption patterns would reflect anomalies in the underlying activity patterns.

In Chapter 2, a feature and model based method for detecting anomalies in activity patterns of an occupant using electrical energy consumption patterns was proposed. The raw energy consumption patterns were believed to be ineffective for reflecting the anomalies in the underlying activity patterns. Therefore, features which were believed to be more effective for reflecting anomalies in activity patterns were extracted from the raw energy consumption data and a model was built based on the features extracted.

The output of the proposed method was a *Score*, a quantitative assessment of anomalies in the energy consumption pattern of a given day.

In Chapter 3, the correlation of the scores provided by the proposed method and the activity patterns of three lone occupants was shown numerically. The proposed feature and model based method was also compared with a chosen raw data based approach. A survey was conducted to obtain evidence to show that anomalies in energy consumption patterns could reflect anomalies in activity patterns. Three lone occupants participated in the survey. The participants were asked to report their hourly activities every day on the survey timesheets for a month. The chosen features for representing the regular (recurring) activity patterns were 1) the daily highly probable activities, 2) daily less probable activities, 3) hourly highly probable activities, 4) hourly less probable activities, 5) daily less probable durations of activities and 6) less probable daily energy consumption.

Using the features of activity patterns, a *Level of Suspicion*, a quantitative assessment of anomalies in a daily activity pattern, could be computed. Fuzzy C-Means clustering (FCM) [44] was used to cluster the level of suspicion values into two clusters (i.e. regular and irregular). The scores provided by the proposed method were also clustered by FCM into two clusters (i.e. regular and irregular). A raw data based approach [32] was chosen to compare with the proposed method. The raw energy consumption sequences were again clustered by FCM and the clusters were classified as regular or irregular.

Using the clustering result of the level of suspicion values as the pseudo ground truth,

the clustering and classification results of the scores and the raw energy consumption sequences were compared. First, the results showed that both the scores and the raw energy sequences correlated with the level of suspicion values. In other words, the results indicated that anomalies in activity patterns can be effectively inferred from energy consumption patterns.

Second, the results showed that the proposed method performs better than the chosen raw data based approach, i.e. extracted features are more effective then the raw energy consumption patterns for reflecting anomalies in the underlying activity patterns.

In Chapter 4, the mRMR feature selection [39] was adopted for reducing the number of features used for representing an energy consumption pattern of a day. The 72 energy features were ranked based on mRMR. It was found that the training set accuracy was improved when the top ranked four features were considered while the test set accuracy remained the same as compared to when all features were considered. It was also found that both training set and test set accuracies were improved when the top ranked 13 features were considered as compared to when all features were considered.

In Chapter 5, we proposed a supervised learning method based on C-ELM (PM-CELM) for classifying the energy consumption patterns as either regular or irregular. PM-CELM requires labelled training data (i.e. a supervised learning). The training data were labelled based on the pseudo ground truth deduced from the survey (introduced in Section 3.5.1). In contrast, the proposed method based on FCM (PM-FCM) introduced in Chapter 2 and Chapter 3 does not require labelled training data (i.e. an unsupervised

learning). It was shown that PM-CELM outperforms PM-FCM. PM-CELM was then enhanced by adopting mRMR feature selection (PM-mRMRnCELM). It was shown that PM-mRMRnCELM can achieve the same accuracy performance as PM-CELM using a subset of the energy features.

The advantage of the proposed method (PM-FCM) is that it detects anomalies in activity patterns of an occupant using household electricity consumption patterns without a need of explicitly monitoring the individual activities of the occupant. Most related works [21–25, 27–30] for detecting anomalies in activities of a person require monitoring the individual activities or actions of the person. Monitoring the individual activities may require a lot more hardware and annotating the individual activities captured by some hardware (e.g. sensors and cameras) could be time-consuming.

The limitations of the proposed method are as follows: 1) it can only detect anomalies in activities that affect energy consumptions, 2) it does not identify which particular activity led to the detected anomalies, and 3) it does not identify whether the detected anomalies indicate a positive or negative change in a person's wellness. However, the proposed method can trigger an alert to caregivers or relatives who can then investigate the cause of the anomalies in a timely manner. If the objective of a certain assisted living system is to automatically raise an alarm when an activity pattern of a monitored person deviates significantly from the regular activity patterns, the proposed method provides a simple solution.

6.2 Future Work

The proposed method (PM-FCM) in this thesis can detect anomalies in activity patterns of an occupant from household energy consumption data, but it does not identify which particular activity led to the anomalies. A possible topic for future research is to recognize individual activities from energy consumption patterns. One will probably need to obtain the activity patterns of an occupant in some way (e.g. survey, sensors) for a relatively long time before one might match a certain activity to a particular energy consumption pattern accurately.

Another limitation of the proposed method in this thesis is that it does not identify the implication of the detected anomalies on a person's wellness. Therefore, another possible topic for future research is to explore the correlation of energy consumption patterns and a person's wellness. A possible intermediate step would be to first investigate the correlation of activity patterns and a person's wellness. One might need to first discover the particular activity pattern (or patterns) which corresponds to a certain change in the person's wellness. One can then observe the corresponding energy consumption patterns every time when that particular activity pattern occurs. After a long term observation, one might be able to relate a certain change in a person's wellness to a particular energy consumption pattern (or patterns).

On the other hand, one might skip the intermediate step mentioned above and directly match a certain change in a person's wellness to a particular energy consumption pattern (or patterns). A major challenge of this future research is to obtain the baseline measure

of a person's wellness. Another difficulty is that it will probably require a large amount of information before one can be sure that a particular energy consumption pattern indicates a certain change in a particular person's wellness.

Bibliography

[1] United Nations, Department of Economic and Social Affairs, Population Division, "World population ageing 2013," 2013.

[2] W. W. Hung, J. S. Ross, K. S. Boockvar, and A. L. Siu, "Recent trends in chronic disease, impairment and disability among older adults in the United States," *BioMedCentral(BMC) Geriatrics*, vol. 11, no. 47, 2011.

[3] Canada, Parliament, House of Commons, Standing Committee on Health, "Evidence," in *Meeting 7, 41st Parliament, 1st Session on*, October 5, 2011.

[4] Canada, Parliament, House of Commons, Standing Committee on Health, "Evidence," in *Meeting 8, 41st Parliament, 1st Session on*, October 17, 2011.

[5] Canada, Parliament, House of Commons, Standing Committee on Health, "Evidence," in *Meeting 12, 41st Parliament, 1st Session on*, October 31, 2011.

[6] R. Harrell, J. Lynott, S. Guzman, and C. Lampkin, "What is livable? community preferences of older adults," *American Association of Retired Persons (AARP) Public Policy Institute*, April 2014.

[7] P. Rashidi and A. Mihailidis, "A survey on ambient-assisted living tools for older adults," *IEEE Journal of Biomedical and Health Informatics*, vol. 17, no. 3, pp. 579–590, 2013.

[8] B. Liu, Y. Zhang, and Z. Liu, "Wearable monitoring system with multiple physiological parameters," in *Medical Devices and Biosensors (MDBS), 5th International Summer School and Symposium on*, 2008, pp. 268–271.

[9] E. Sardini, M. Serpelloni, and M. Ometto, "Multi-parameters wireless shirt for physiological monitoring," in *Medical Measurements and Applications Proceedings (MeMeA), IEEE International Workshop on*, 2011, pp. 316–321.

[10] K. Malhi, S. C. Mukhopadhyay, J. Schnepper, M. Haefke, and H. Ewald, "A zigbee-based wearable physiological parameters monitoring system," *IEEE Sensors Journal*, vol. 12, no. 3, pp. 423–430, 2012.

[11] R. Logier *et al.*, "A multi sensing method for robust measurement of physiological parameters in wearable devices," in *Engineering in Medicine and Biology Society (EMBS), 36th Annual International Conference of the IEEE on*, 2014, pp. 994–997.

[12] J. Chen, K. Kwong, D. Chang, J. Luk, and R. Bajcsy, "Wearable sensors for reliable fall detection," in *Engineering in Medicine and Biology Society (EMBS), 27th Annual International Conference of the IEEE on*, 2006, pp. 3551–3554.

[13] T. Tamrat, M. Griffin, S. Rupcic, S. Kachnowski, T. Taylor, and J. Barfield, "Operationalizing a wireless wearable fall detection sensor for older adults," in *Pervasive*

Computing Technologies for Healthcare (PervasiveHealth), 6th International Conference of the IEEE on, 2012, pp. 297–302.

[14] M. Popescu, Y. Li, M. Skubic, and M. Rantz, "An acoustic fall detector system that uses sound height information to reduce the false alarm rate," in *Engineering in Medicine and Biology Society (EMBS), 30th Annual International Conference of the IEEE on*, 2008, pp. 4628–4631.

[15] X. Zhuang, J. Huang, G. Potamianos, and M. Hasegawa-Johnson, "Acoustic fall detection using gaussian mixture models and GMM supervectors," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP) of the IEEE on*, 2009, pp. 69–72.

[16] C. Rougier, J. Meunier, A. St-Arnaud, and J. Rousseau, "Robust video surveillance for fall detection based on human shape deformation," *Circuits and Systems for Video Technology, IEEE Transactions on*, vol. 21, no. 5, pp. 611–622, 2011.

[17] V. Vaidehi, K. Ganapathy, K. Mohan, A. Aldrin, and K. Nirmal, "Video based automatic fall detection in indoor environment," in *International Conference on Recent Trends in Information Technology (ICRTIT) of the IEEE on*, 2011, pp. 1016–1020.

[18] Z. Zhou, X. Chen, Y.-C. Chung, Z. He, T. X. Han, and J. M. Keller, "Activity analysis, summarization, and visualization for indoor human activity monitoring,"

Circuits and Systems for Video Technology, IEEE Transactions on, vol. 18, no. 11, pp. 1489–1498, 2008.

[19] L. Chen, J. Hoey, C. D. Nugent, D. J. Cook, and Z. Yu, "Sensor-based activity recognition," *Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on*, vol. 42, no. 6, pp. 790–808, 2012.

[20] O. D. Lara and M. A. Labrador, "A survey on human activity recognition using wearable sensors," *IEEE Communications Surveys & Tutorials*, vol. 15, no. 3, pp. 1192–1209, 2013.

[21] N. Suryadevara, A. Gaddam, S. Mukhopadhyay, and R. Rayudu, "Wellness determination of inhabitant based on daily activity behaviour in real-time monitoring using sensor networks," in *Sensing Technology (ICST), 5th International Conference of the IEEE on*, 2011, pp. 474–481.

[22] N. K. Suryadevara and S. C. Mukhopadhyay, "Wireless sensor network based home monitoring system for wellness determination of elderly," *IEEE Sensors Journal*, vol. 12, no. 6, pp. 1965–1972, 2012.

[23] N. Suryadevara, S. Mukhopadhyay, R. Wang, R. K. Rayudu, and Y. Huang, "Reliable measurement of wireless sensor network data for forecasting wellness of elderly at smart home," in *Instrumentation and Measurement Technology Conference (I2MTC) of the IEEE on*, 2013, pp. 16–21.

[24] N. Suryadevara and S. Mukhopadhyay, "An intelligent system for continuous monitoring of wellness of an inhabitant for sustainable future," in *Humanitarian Technology Conference (R10-HTC) of the IEEE on*, 2014, pp. 70–75.

[25] N. K. Suryadevara and S. C. Mukhopadhyay, "Determining wellness through an ambient assisted living environment," *IEEE Intelligent Systems*, vol. 29, no. 3, pp. 30–37, 2014.

[26] B. O'Mullane, B. Bortz, A. O'Hannlon, J. Loane, and R. B. Knapp, "Comparison of health measures to movement data in aware homes," *Ambient Intelligence, Springer*, pp. 290–294, 2011.

[27] V. Jakkula, D. J. Cook, *et al.*, "Anomaly detection using temporal data mining in a smart home environment," *Methods of Information in Medicine*, vol. 47, no. 1, pp. 70–75, 2008.

[28] K.-J. Kim, M. M. Hassan, S. Na, and E.-N. Huh, "Dementia wandering detection and activity recognition algorithm using tri-axial accelerometer sensors," in *Ubiquitous Information Technologies & Applications (ICUT), 4th International Conference of the IEEE on*, 2009, pp. 1–5.

[29] C. Franco, J. Demongeot, C. Villemazet, and N. Vuillerme, "Behavioral telemonitoring of the elderly at home: Detection of nycthemeral rhythms drifts from location data," in *Advanced Information Networking and Applications Workshops (WAINA), 24th International Conference of the IEEE on*, 2010, pp. 759–766.

[30] E. Campo, M. Chan, W. Bourennane, and D. Estève, "Behaviour monitoring of the elderly by trajectories analysis," in *Engineering in Medicine and Biology Society (EMBS), Annual International Conference of the IEEE on*, 2010, pp. 2230–2233.

[31] T. W. Liao, "Clustering of time series data – a survey," *Pattern Recognition*, vol. 38, no. 11, pp. 1857–1874, 2005.

[32] F. Iglesias and W. Kastner, "Analysis of similarity measures in times series clustering for the discovery of building energy patterns," *Energies*, vol. 6, no. 2, pp. 579–597, 2013.

[33] P. C. Mahalanobis, "On the generalized distance in statistics," in *Proceedings of the National Institute of Sciences*, vol. 2, 1936, pp. 49–55.

[34] D. J. Berndt and J. Clifford, "Using dynamic time warping to find patterns in time series," in *Association for the Advancement of Artificial Intelligence (AAAI) Workshop on Knowledge Discovery in Databases on*, 1994, pp. 359–370.

[35] K. Pearson, "Mathematical contributions to the theory of evolution –on a form of spurious correlation which may arise when indices are used in the measurement of organs," in *Proceedings of the Royal Society of London*, vol. 60, 1896, pp. 489–498.

[36] J. C. Dunn, "A fuzzy relative of the isodata process and its use in detecting compact well-separated clusters," *Journal of Cybernetics*, pp. 32–57, 1974.

[37] D. L. Davies and D. W. Bouldin, "A cluster separation measure," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, no. 2, pp. 224–227, 1979.

[38] Y. Jung, H. Park, D.-Z. Du, and B. L. Drake, "A decision criterion for the optimal number of clusters in hierarchical clustering," *Journal of Global Optimization*, vol. 25, no. 1, pp. 91–111, 2003.

[39] H. Peng, F. Long, and C. Ding, "Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 27, no. 8, pp. 1226–1238, 2005.

[40] C. Ding and H. Peng, "Minimum redundancy feature selection from microarray gene expression data," *Journal of bioinformatics and computational biology*, vol. 3, no. 2, pp. 185–205, 2005.

[41] Q. Wu and C. Miao, "C-ELM: A curious extreme learning machine for classification problems," in *Proceedings of ELM-2014*, vol. 1, pp. 355–366.

[42] BC Hydro. Smart metering program. Accessed: 2015-06-05. [Online]. Available: https://www.bchydro.com/energy-in-bc/projects/smart_metering_infrastructure_program.html

[43] E. Kreyszig, *Advanced Engineering Mathematics Tenth Edition*. Wiley, 2011, p. 1014.

[44] J. C. Bezdek, R. Ehrlich, and W. Full, "FCM: The fuzzy c-means clustering algorithm," *Computers & Geosciences*, vol. 10, no. 2, pp. 191–203, 1984.

[45] U. Maulik and S. Bandyopadhyay, "Performance evaluation of some clustering algorithms and validity indices," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 24, no. 12, pp. 1650–1654, 2002.

[46] H. Liu, J. Li, and L. Wong, "A comparative study on feature selection and classification methods using gene expression profiles and proteomic patterns," *Genome Informatics*, vol. 13, pp. 51–60, 2002.

[47] S. Doraisamy, S. Golzari, N. Mohd, M. N. Sulaiman, and N. I. Udzir, "A study on feature selection and classification techniques for automatic genre classification of traditional Malay music." in *the 9th International Conference on Music Information Retrieval on*, 2008, pp. 331–336.

[48] G. Forman, "An extensive empirical study of feature selection metrics for text classification," *The Journal of Machine Learning Research*, vol. 3, pp. 1289–1305, 2003.

[49] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," *The Journal of Machine Learning Research*, vol. 3, pp. 1157–1182, 2003.

[50] K. Kira and L. A. Rendell, "The feature selection problem: traditional methods and a new algorithm," in *the 10th national conference on Artificial intelligence on*, 1992, pp. 129–134.

[51] C. Cortes and V. Vapnik, "Support-vector networks," *Machine Learning*, vol. 20, no. 3, pp. 273–297, 1995.

[52] R. Kohavi and G. H. John, "Wrappers for feature subset selection," *Artificial Intelligence*, vol. 97, no. 1, pp. 273–324, 1997.

[53] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik, "Gene selection for cancer classification using support vector machines," *Machine Learning*, vol. 46, no. 1-3, pp. 389–422, 2002.

[54] E. Frank, M. Hall, and B. Pfahringer, "Locally weighted naive Bayes," in *the 19th conference on Uncertainty in Artificial Intelligence on*, 2002, pp. 249–256.

[55] J. Liu, S. Ranka, and T. Kahveci, "Classification and feature selection algorithms for multi-class CGH data," *Bioinformatics*, vol. 24, no. 13, pp. i86–i95, 2008.

[56] A. Janecek, W. N. Gansterer, M. Demel, and G. Ecker, "On the relationship between feature selection and classification accuracy," in *Journal of Machine Learning Research (JMLR), Workshop and Conference on*, vol. 4, 2008, pp. 90–105.

Appendices

Appendix A

Pseudocode for the Proposed Method

Algorithm 6 The pseudocode for the proposed method

Let D be the number of days in history
int[ ][ ] $maxTimeCount$ = new int [24][24]
int[ ][ ] $minTimeCount$ = new int [24][24]
double[ ][ ] $range$ = new double [24][D]
for $d = 1$ to D **do**
  for $i = 1$ to 24 **do**
    Get the i-hour moving total sequence of day d
    $maxTime$ = time of day when the Maximum occurred
    $minTime$ = time of day when the Minimum occurred
    $range[i][d]$ = Maximum - Minimum
    $maxTimeCount[i][maxTime]$ increased by 1
    $minTimeCount[i][minTime]$ increased by 1
  end for
end for
*** Deduce regular energy patterns from history ****
double[ ][ ] $probMax$ = new double [24][24]
double[ ][ ] $probMin$ = new double [24][24]
$probMax = maxTimeCount/D$
$probMin = minTimeCount/D$
double[ ] $rangeMean$ = new double [24]
double[ ] $rangeStdev$ = new double [24]
for $i = 1$ to 24 **do**
  $rangeMean[i]$ = Mean($range[i][1{:}D]$)
  $rangeStdev[i]$ = Standard Deviation($range[i][1{:}D]$)
end for

Algorithm 6 The pseudocode for the proposed method (continued)

```
*** Calculate the score for a given day ****
*** Assuming the score assignments are +/- 1 ****
int[ ] scoreMax = new int [24]
int[ ] scoreMin = new int [24]
int[ ] scoreRange = new int [24]
for i = 1 to 24 do
  Get the i-hour moving total sequence of the day of interest
  thisMaxTime = time of day when the Maximum occurred
  thisMinTime = time of day when the Minimum occurred
  thisRange = Maximum - Minimum
  thisRangeZscore = abs(thisRange - rangeMean[i])/rangeStdev[i]
  if probMax[i][thisMaxTime] ≥ thresholdProbMax then
    scoreMax[i] = 1
  else
    scoreMax[i] = -1
  end if
  if probMin[i][thisMinTime] ≥ thresholdProbMin then
    scoreMin[i] = 1
  else
    scoreMin[i] = -1
  end if
  if thisRangeZscore ≤ thresholdRangeZscore then
    scoreRange[i] = 1
  else
    scoreRange[i] = -1
  end if
end for
FinalScore = sum(scoreMax[1:24])+sum(scoreMin[1:24])+sum(scoreRange[1:24])
```

Appendix B

Survey Samples from Home C

The survey timesheets of March 2 and March 15 from Home C are provided in this appendix.

Name: Home C

Date (mm/dd/yyyy): 03/02/2015

This excel sheet is intended to help you keep track of your activities of a day. For each hour (each column), please check all activities that apply and all appliances that you used during the hour.

	00:01 - 01:00	01:01 - 02:00	02:01 - 03:00	03:01 - 04:00	04:01 - 05:00	05:01 - 06:00	06:01 - 07:00	07:01 - 08:00
Not at home								
Sleeping	X	X	X	X	X	X	X	
Bathing								
Dining at home								
Electric cooking appliances (oven, rice cooker, etc.)								X
Dishwasher								
House cleaning (vacuum cleaner, etc.)								
Entertainment (TV, computer, radio, etc.)								X
Laundry (washer, dryer, iron, etc.)								
Heater / Air conditioner							X	X
Other energy-consuming activities/appliances* (see note at the bottom of the page)								

	08:01 - 09:00	09:01 - 10:00	10:01 - 11:00	11:01 - 12:00	12:01 - 13:00	13:01 - 14:00	14:01 - 15:00	15:01 - 16:00
Not at home			X	X	X	X	X	X
Sleeping								
Bathing								
Dining at home		X						
Electric cooking appliances (oven, rice cooker, etc.)	X	X						
Dishwasher								
House cleaning (vacuum cleaner, etc.)								
Entertainment (TV, computer, radio, etc.)	X	X						
Laundry (washer, dryer, iron, etc.)								
Heater / Air conditioner	X							
Other energy-consuming activities/appliances* (see note at the bottom of the page)								

	16:01 - 17:00	17:01 - 18:00	18:01 - 19:00	19:01 - 20:00	20:01 - 21:00	21:01 - 22:00	22:01 - 23:00	23:01 - 00:00
Not at home	X	X	X	X	X			
Sleeping								X
Bathing							X	
Dining at home							X	
Electric cooking appliances (oven, rice cooker, etc.)							X	
Dishwasher								
House cleaning (vacuum cleaner, etc.)								
Entertainment (TV, computer, radio, etc.)						X	X	X
Laundry (washer, dryer, iron, etc.)								
Heater / Air conditioner								
Other energy-consuming activities/appliances* (see note at the bottom of the page)								

**Activities/appliances that you think would consume a lot of power except for those always-on appliances (fridge, light, etc.)*

Figure B.1: Home C survey timesheet - Mar 2,2015

Name: Home C

Date (mm/dd/yyyy): 03/15/2015

This excel sheet is intended to help you keep track of your activities of a day. For each hour (each column), please check all activities that apply and all appliances that you used during the hour.

	00:01 - 01:00	01:01 - 02:00	02:01 - 03:00	03:01 - 04:00	04:01 - 05:00	05:01 - 06:00	06:01 - 07:00	07:01 - 08:00
Not at home								
Sleeping	X	X	X	X	X	X	X	
Bathing								
Dining at home								
Electric cooking appliances (oven, rice cooker, etc.)								X
Dishwasher								
House cleaning (vacuum cleaner, etc.)								
Entertainment (TV, computer, radio, etc.)								X
Laundry (washer, dryer, iron, etc.)								
Heater / Air conditioner							X	X
Other energy-consuming activities/appliances* (see note at the bottom of the page)								

	08:01 - 09:00	09:01 - 10:00	10:01 - 11:00	11:01 - 12:00	12:01 - 13:00	13:01 - 14:00	14:01 - 15:00	15:01 - 16:00
Not at home			X	X	X	X	X	X
Sleeping								
Bathing								
Dining at home		X						
Electric cooking appliances (oven, rice cooker, etc.)	X	X						
Dishwasher								
House cleaning (vacuum cleaner, etc.)								
Entertainment (TV, computer, radio, etc.)	X	X						
Laundry (washer, dryer, iron, etc.)								
Heater / Air conditioner	X							
Other energy-consuming activities/appliances* (see note at the bottom of the page)								

	16:01 - 17:00	17:01 - 18:00	18:01 - 19:00	19:01 - 20:00	20:01 - 21:00	21:01 - 22:00	22:01 - 23:00	23:01 - 00:00
Not at home	X	X	X	X	X			
Sleeping								X
Bathing						X		
Dining at home								
Electric cooking appliances (oven, rice cooker, etc.)							X	
Dishwasher								
House cleaning (vacuum cleaner, etc.)								
Entertainment (TV, computer, radio, etc.)						X	X	X
Laundry (washer, dryer, iron, etc.)								
Heater / Air conditioner								
Other energy-consuming activities/appliances* (see note at the bottom of the page)								

**Activities/appliances that you think would consume a lot of power except for those always-on appliances (fridge, light, etc.)*

Figure B.2: Home C survey timesheet - Mar 15,2015

Appendix C

Choosing the Design Variable Values for the Proposed Method

This appendix explains how the design variables were chosen using Home B as an example. Please refer to Table 3.1 in Section 3.3.1 for the list of the chosen thresholds of Home B. This appendix explains how the thresholds were chosen according to the hourly energy consumption data. Since the two to 24 hours moving totals were essentially derived from the hourly energy consumption data, the same set of thresholds was applied to all of them.

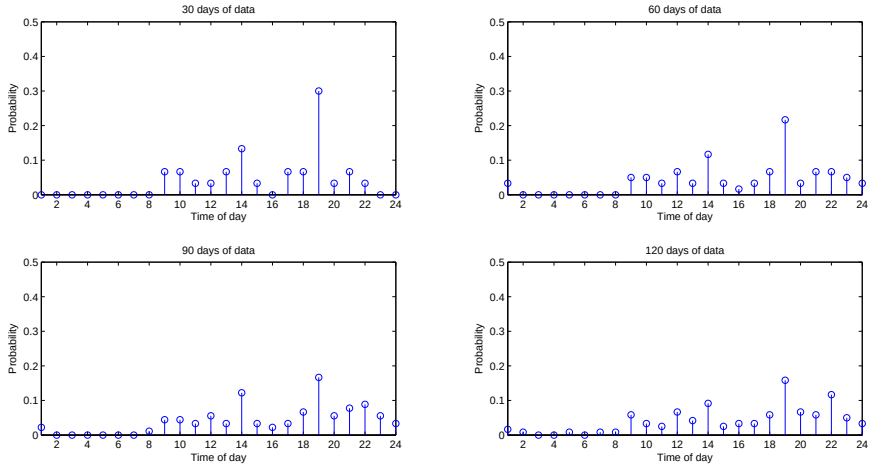

Figure C.1: Occurrence probability of the maximum hourly energy consumption at Home B considering 30, 60, 90 and 120 days of data respectively

First, the number of days of data to be included was determined according to the

hourly energy consumption data. As can be seen in Figure C.1, the maximum hourly energy consumption occurred more frequently at 2PM and 7PM and this was the most evident when only 30 days of data were included. As more data were included, the occurrence probabilities of the maximum hourly energy consumption at 2PM and 7PM tended to distribute to adjacent hours. This is probably because the activity patterns of the occupant change slightly over time. On the one hand, we would like to capture only the most recent regular activity patterns of the occupant. On the other hand, we would also like to include as much data as possible so that we have more confidence in the probability estimation. Considering the trade-off between the above two objectives, the size of the data window was chosen to be 60 days.

Second, the threshold between the probable and less probable hours needed to be determined. The premise was that we would rather tolerate more false alarms than more missed detections of anomalies in activity patterns. Therefore, the threshold value was chosen so that there were a lot more less probable hours than probable hours. In other words, only the hours with significantly higher probabilities than the others were deemed probable hours. As illustrated in the top right sub-plots of Figure C.1, there were only two noticeable peaks (at 2PM and 7PM). The threshold was then chosen so that only 2PM and 7PM were deemed probable. A probability of 0.1 appeared to be a reasonable threshold to achieve this. The threshold value for the occurrence of the minimum hourly energy consumption was chosen in a similar manner and hence is not discussed in details in this appendix.

Third, the threshold value for the range Z-Score was determined. The threshold value was again determined according to the hourly energy consumption data. After observing several different 60-day data windows, the average of the absolute range Z-Scores of the hourly energy consumption data was about 0.8. However, if the average was used as the threshold, there would be too many days deemed not normal from a range Z-Score's perspective. Therefore, a threshold higher than the average was more desirable. The threshold value was then chosen to be 1 which is about 25% higher than the average.

Last, the score assignment was determined. As there was no particular reason to weight one feature more than the others, all the sub-scores were equally weighed. Therefore, the probabilities on or above the corresponding thresholds were assigned +1 while those below the thresholds were assigned -1. The range Z-Scores on or below the threshold were assigned +1 while those above the threshold were assigned -1.

Appendix D

Choosing the Threshold Values of Regular Activity Patterns for the Survey

This appendix explains how the threshold values were chosen, using Home B as an example. Please refer to Table 3.2 in Section 3.3.2 for the list of the chosen threshold values and Table 3.3 in Section 3.3.2 for the length of data included for Home B.

First, the number of days of daily energy consumption data (i.e. the total energy consumption of all appliances during a day) to be included was determined. As the daily energy consumption generally did not vary much from day to day, 365 days of the daily energy consumption data were included in case there were any seasonal changes. Although the daily energy consumption data did not vary much from day to day, the hourly energy consumption data within a day fluctuated noticeably from day to day. Therefore, the size of the data window of the hourly energy consumption data mentioned in Appendix C was much smaller.

Second, the threshold values of regular activity patterns were determined. The best scenario was to have a few days with noticeably higher level of suspicion values and to have most of the other days with lower level of suspicion values. In order to achieve this, the combinations of the six threshold values needed to be considered. A list of the

Features of Activity Patterns	Potential Thresholds
	Probability
Daily highly probable activities	$\geq 0.8/0.9$
Daily less probable activities	$\leq 0.2/0.1$
Hourly highly probable activities	$\geq 0.8/0.85/0.9/0.95$
Hourly less probable activities	$\leq 0.15/0.1/0.05$
Daily less probable durations of activities	$\leq 0.15/0.1/0.05$
	|Z-Score|
Less probable daily energy consumptions	$\geq 0.9/0.95/1/1.2/1.3$

Table D.1: A list of potential threshold values for Home B

potential thresholds for Home B are shown in Table D.1. The potential thresholds were chosen by observation. For instance, 0.7 was not included in the list of the daily highly probable activities because there was no activity that occurred between 43% and 80% of the total number of days.

As mentioned in Section 3.2.7, the accumulated level of suspicion value for a day is equal to the sum of the level of suspicion values from the above six features. If the activity pattern of a certain day is significantly different from the regular activity patterns, the accumulated level of suspicion value of that particular day will probably be due to most of the six features. If the level of suspicion value of a day is relatively low and is only due to one feature, it probably indicates that the threshold value for that particular feature can be tightened. Therefore, the selection of the threshold values began with the loosest values for the six features. The next threshold values were chosen when that helped differentiate the days with noticeably higher level of suspicion values from the other days. The selection of threshold values ended when one of the following events was encountered: 1) the selection of the thresholds was already the tightest possible

combination and 2) choosing the next threshold value of any of the six features would significantly harm the differentiation between the highly suspicious (possibly anomalous) days and the possibly normal days.

Appendix E

Performance Evaluation without Considering Daily Energy Consumption

	Proposed Method	Raw Data Based Approach [32]
Training Set		
Accuracy	46/61 (75.41%)	43/61 (70.49%)
Missed Detection	5	4
False Alarm	10	14
Test Set		
Accuracy	22/30 (73.33%)	19/30 (63.33%)
Missed Detection	4	4
False Alarm	4	7

Table E.1: The performances of the proposed method and the chosen raw data based approach without considering daily energy consumption

The pseudo ground truth in this evaluation was derived from a modified version of the level of suspicion value. The modified level of suspicion value only considers the five features described in Section 3.2.1 to Section 3.2.5, but does not include the daily energy consumption feature described in Section 3.2.6. The purpose of this evaluation was to examine the impact of the daily energy consumption feature on the accuracy performances of the proposed method and the raw data based approach. The performances of the proposed method and the raw data based approach are summarized in Table E.1. As compared to the accuracy performances in Section 3.5.5, the training set accuracies of the proposed method and the raw data based approach dropped by 19.30% ($\frac{93.44-75.41}{93.44} \times 100\%$)

and 20.37% respectively, while the test set accuracies dropped by 15.38% and 13.64% respectively. Removing the less probable daily energy consumption feature from the pseudo ground truth appeared to adversely affect both the proposed method and the raw data based approach equally.

Appendix F

Rankings of the Energy Features for Homes A, B and C

The rankings of the 72 energy features for Homes A, B and C based on mRMR feature selection (with the MIQ criterion) are shown in this appendix. A dictionary of the feature indices is also provided.

Rank	Home A	Home B	Home C	Rank	Home A	Home B	Home C
	Feature Index				Feature Index		
1	65	52	63	37	4	25	7
2	33	50	66	38	60	9	38
3	9	44	61	39	62	59	42
4	51	49	51	40	14	48	28
5	69	66	57	41	27	60	17
6	16	26	64	42	29	11	41
7	23	31	70	43	12	61	47
8	64	53	65	44	22	62	14
9	6	51	58	45	20	17	46
10	52	67	60	46	30	36	8
11	67	39	52	47	7	63	5
12	54	68	53	48	8	7	48
13	32	54	69	49	5	8	29
14	66	69	56	50	45	32	21
15	53	46	72	51	43	37	25
16	31	27	54	52	3	14	37
17	55	43	62	53	47	6	31
18	70	70	55	54	25	16	27
19	68	55	67	55	46	29	26
20	17	71	49	56	24	12	43
21	56	56	59	57	15	20	19
22	10	72	68	58	36	15	44
23	57	41	50	59	2	33	15
24	71	34	71	60	38	47	1
25	49	57	32	61	42	30	36
26	58	40	40	62	1	13	23
27	72	10	22	63	18	23	2
28	50	18	34	64	40	22	4
29	28	35	35	65	37	38	18
30	59	3	6	66	19	1	24
31	11	64	3	67	26	21	39
32	61	65	16	68	41	4	13
33	21	19	11	69	48	42	9
34	63	2	30	70	35	5	20
35	34	28	45	71	44	24	10
36	13	58	33	72	39	45	12

Table F.1: The rankings of the energy features for Homes A, B and C based on mRMR with the MIQ criterion

Index	Feature	Index	Feature
1	1- hour Max time	37	13- hour Min time
2	2- hour Max time	38	14- hour Min time
3	3- hour Max time	39	15- hour Min time
4	4- hour Max time	40	16- hour Min time
5	5- hour Max time	41	17- hour Min time
6	6- hour Max time	42	18- hour Min time
7	7- hour Max time	43	19- hour Min time
8	8- hour Max time	44	20- hour Min time
9	9- hour Max time	45	21- hour Min time
10	10- hour Max time	46	22- hour Min time
11	11- hour Max time	47	23- hour Min time
12	12- hour Max time	48	24- hour Min time
13	13- hour Max time	49	1- hour Range
14	14- hour Max time	50	2- hour Range
15	15- hour Max time	51	3- hour Range
16	16- hour Max time	52	4- hour Range
17	17- hour Max time	53	5- hour Range
18	18- hour Max time	54	6- hour Range
19	19- hour Max time	55	7- hour Range
20	20- hour Max time	56	8- hour Range
21	21- hour Max time	57	9- hour Range
22	22- hour Max time	58	10- hour Range
23	23- hour Max time	59	11- hour Range
24	24- hour Max time	60	12- hour Range
25	1- hour Min time	61	13- hour Range
26	2- hour Min time	62	14- hour Range
27	3- hour Min time	63	15- hour Range
28	4- hour Min time	64	16- hour Range
29	5- hour Min time	65	17- hour Range
30	6- hour Min time	66	18- hour Range
31	7- hour Min time	67	19- hour Range
32	8- hour Min time	68	20- hour Range
33	9- hour Min time	69	21- hour Range
34	10- hour Min time	70	22- hour Range
35	11- hour Min time	71	23- hour Range
36	12- hour Min time	72	24- hour Range

Table F.2: Dictionary of the feature indices