# INVESTIGATING THE HUMAN BEHAVIOR SIDE OF BUILDING ENERGY EFFICIENCY

by

CHAO CHEN

A dissertation submitted in partial fulfillment of
the requirements for the degree of

DOCTOR OF PHILOSOPHY

WASHINGTON STATE UNIVERSITY
Department of Electrical Engineering and Computer Science

AUGUST 2013

To the Faculty of Washington State University:

The members of the Committee appointed to examine the dissertation of CHAO CHEN find it satisfactory and recommend that it be accepted.

Diane J. Cook, Ph.D., Chair

Lawrence B. Holder, Ph.D.

Behrooz Shirazi, Ph.D.

# INVESTIGATING THE HUMAN BEHAVIOR SIDE OF BUILDING ENERGY EFFICIENCY

## Abstract

by Chao Chen, Ph.D.
Washington State University
August 2013

Chair: Diane J. Cook

Society is becoming increasingly aware of the impact that our lifestyle preferences have on energy usage and the environment. In this dissertation, we look more closely at the impact that human behavior has on energy consumption. In particular, we design and evaluate smart home and machine learning techniques to examine the relationship between behavioral patterns and resource consumption.

The contribution of this research has two components. In the first part, we use smart home technologies to examine the relationship between behavior patterns and energy usage at the scale of individual homes. In particular, machine learning techniques are used to predict energy usage based on residents activities. Data mining techniques are then introduced to identify anomalies and abnormal patterns in home power data. Lastly, a web-based tool for visualizing smart home activities and power

consumption is designed. This tool is used to present the results of the previous techniques to inform users about their personal energy usage and to encourage more energy-efficient behaviors.

In the second part, we focus on the creation of data mining algorithms to analyze per-home energy use at the scale of an entire community. Specifically, we analyze data collected from thousands of smart meters. Our contributions to such large-scale analysis include the design of an automated outlier detection tool for noise reduction. In addition, a web-based visualization system was developed to depict an energy usage heat map and compare residential electricity usage between neighboring homes. In order to build a more complete model of electricity usage, we designed a learning algorithm that takes into account the function of various building characteristics. Using this model, a web-based user interface was designed to estimate building energy usage. Finally, an unsupervised algorithm was designed to cluster large-scale time-series datasets in an efficient manner.

We describe and evaluate each of these contributions using electricity consumption data from actual smart homes as part of the CASAS smart home project. In each case we illustrate the efficacy of these algorithms to gaining insights on human behavior and its impact on energy consumption, and offer ideas for using these insights to promote sustainable behaviors.

# TABLE OF CONTENTS

# LIST OF TABLES

# LIST OF FIGURES

# CHAPTER 1. INTRODUCTION

In 2011, the United States consumed 97,301 quadrillion btu of power energy. This power consumption is a 200% increase from 1949 [U.S. Energy Information Administration, 2012a]. In recent decades, energy consumption has been outpacing power production, which places a strong demand upon US energy imports, as shown in Figure 1.1. The growth of energy usage is not entirely due to manufacturing plants, industrial production, and automobiles, as is often assumed. In fact, the residential sector is responsible for 16-50% of energy consumption consumed by all sectors worldwide [Ichinose et al., 1999]. In the United States, buildings consumed approximately 74% of the electricity and 34% of the natural gas. Thus, energy consumption of buildings is responsible for 40% of the carbon dioxide emissions [U.S. Energy Information Administration, 2012b]. Society is becoming increasingly aware of the impact residential lifestyle choices make on energy usage and the environment. As a result, there is an urgent need to develop technologies that examine energy usage in homes and buildings to encourage energy efficient behaviors.

Researchers have shown that energy expenditure can be reduced by 5-15% in homes just as a response to acquiring and viewing raw energy usage [Darby, 2006]. However, most residents still receive little or no detailed feedback about their per-

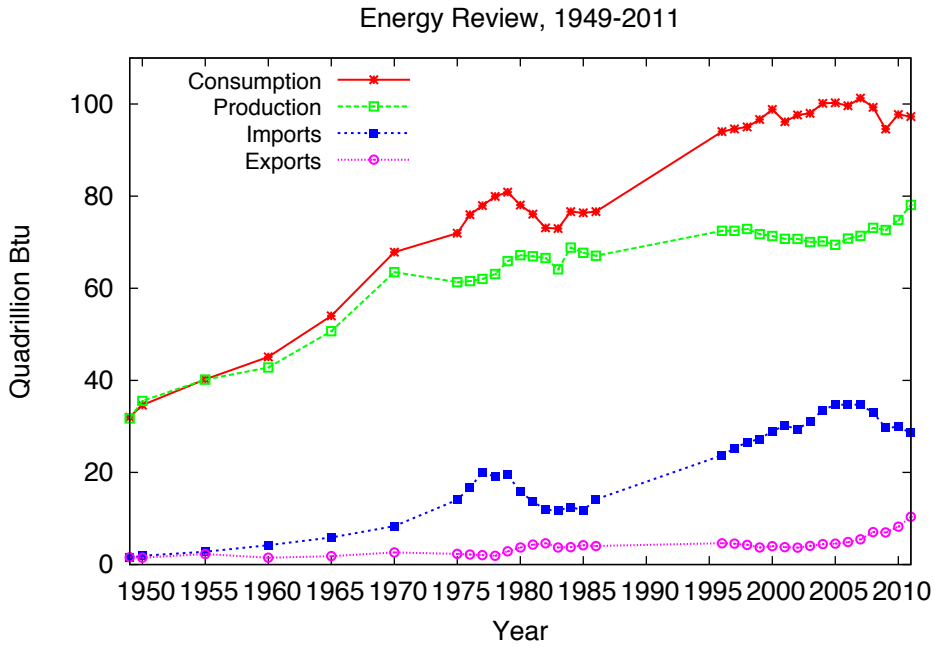



**Figure 1.1:** Primary energy overview during 1949-2011 [U.S. Energy Information Administration, 2012a].

sonal energy usage. A typical utility bill provides information only about monthly energy consumption and a total price to be paid, leaving residents to guess what might explain a higher or lower bill. Earlier studies indicate that residential behavior can influence energy usage by as much as 100% in a given house [Seryak and Kissock, 2003]. Thus, behavior-based energy information is capable of encouraging individuals to modify habits in energy-efficient ways that would be beneficial for both the household and community. However, occupants' behaviors are difficult to capture

and self-report of behavior is typically unreliable and prone to error [Szewcyzk et al., 2009].

A smart home environment [Cook and Das, 2004] is one that acquires and applies knowledge about its residents and their physical surroundings in order to improve their experience in that setting. Such home environments, equipped with sensors for detecting motion, light level, temperature, and energy and water consumption, are ideal testbeds for investigating techniques of inducing behavioral changes to reduce energy usage. We hypothesize that resident behaviors will exercise a great influence on energy consumption in homes. This view is supported by an increasing body of work that links awareness of energy consumption and its impact to behavioral change [Darby, 2010]. By associating activities with energy use and costs, intelligent systems can be devised to automatically control home environments so as to improve energy efficiency and cut expenses.

We investigate the connection between behavior and energy consumption from two points of view. First, we monitor activities as well as energy consumption in an individual smart home to explicitly model the correlation between the two parameters. Second, we examine a larger-scale data collection in order to identify patterns in energy consumption at a community level. For the first part of our work, we propose to use smart home concepts and technologies to provide these important insights. The long-term vision for this work is to enhance understanding of human resource

consumption and to provide tools that promote resource efficiency in smart homes. We hypothesize that energy consumption is correlated with the type of activities that are performed and can therefore be predicted based on automatically-recognized activities that occur in a smart environment. To validate this hypothesis, classification and regression models are built to predict energy consumption in the home as a function of sensor events and time features. Additionally, we further postulate that patterns and anomalies can be automatically detected from energy consumption data and that these discoveries can provide insights on behavioral patterns. Traditional statistically-based detection algorithms can only identify extreme individual power data points. In contrast, structural power patterns may be more important for the end-users. To detect such anomalies, we introduce a pattern mining technique to analyze numeric energy data after mapping numeric values to symbols. These hypotheses are validated by designing algorithms to mine smart home data and evaluating the algorithms using data collected in real world smart home testbeds. We also hypothesize that users need to be provided feedback to motivate behavior change for reducing energy usage and that the feedback can be provided by email, smart phones, or other digital communication mediums. To implement this hypothesis, a web-based visualizer is developed for visualizing raw sensor events and suggesting more energy-efficient behaviors to users.

In the second part of our work, large-scale power monitoring for one year in

the Pullman metropolitan area was examined. We first hypothesize that the outliers existing in raw data are detectable and important to mitigate. To replace these outliers with reasonable values, several regression models are developed for detecting and simulating approximated replacement values. In addition, we hypothesize that the modeling of building energy consumption can be a helpful guide for economic and political decisions regarding energy demand and supply, resource optimization, and building design. In our work, energy consumption models of large-scale buildings are used to quantify energy requirements based on historic energy profiles and building features available from public tax assessor records. To utilize our algorithms, two web-based visualization systems are designed. The first tool is targeted for utilities, as a method to visualize heat maps of energy usage in the whole city, allowing companies to identify abnormal customers quickly. The second tool, targeted to end-users, determines their possible energy consumption, and provides alarms when their energy profile varies greatly from similar buildings.

Finally, we design a method of comparing building-level energy usage with usage by geographic or similarity-based neighbors. We suggest that such information can be provided to residents in order to reduce energy usage by comparing individuals with their neighbours. Researchers [Allcott, 2011] also noted that information on social norms can encourage occupants to significantly conserve energy by comparing their own energy usage with their neighbours. We first split long-term power data

into several small periods then for each period group together residents with similar power profiles. We next use a similarity factor to estimate the similarity between residents and finally identify similar users by measuring the level of similarity. All of our algorithms and techniques are described and evaluated using real data collected in actual homes around the Pullman, Washington area.

# CHAPTER 2. RELATED WORKS

Energy modeling, management, and anomaly detection in building environments are all challenging research problems that have been studied for several decades. Solutions to these problems often rely upon a mixture of different disciplines, including power engineering, statistics, machine learning, and network management. This chapter summarizes existing research and industry efforts in energy-efficient fields on modeling, identifying and managing building energy consumption.

## 2.1 Smart Home Projects

Due to the increasing need to support individuals living in homes independently and comfortably with pervasive computing technologies, a number of approaches and frameworks are designed to implement smart homes that are being applied to a wide range of health, family entertainment, and sustainability applications. Here we describe a few representative examples that not only reflect research ideas of the creators but have been fully implemented and realized.

While many groups have discussed possible approaches to design smart homes, a smaller set of projects describe the system architecture that has been used in actual

smart home implementations. As an example, the Gator Tech Smart House [Helal et al., 2005] provides a clear hierarchical structure with five functional layers: physical layer, sensor layer, service layer, context management layer and application layer. Each layer is responsible for its own function. This type of design can be easily applied to a variety of houses and buildings with different spaces. Based on a similar design philosophy, the MavHome project [Cook et al., 2003] implements an intelligent agent that perceives real-time home states, and then acts upon the home by device controllers. The purpose of this project is to maximize inhabitant comfort and minimize the overall energy cost in the home. For a goal of sustainable living, the Duke Smart Home [Duke Smart Home, 2013] provides a general research and education platform to encourage more students to explore energy-efficient home designs and research projects. To deal with the aging problem, the Aware Home [Aware Home Research Initiative, 2013] focuses on helping elderly residents stay in their own home independently and reducing the burden of caregivers. To meet the increasing needs of the entertainment, the MIT House_n [MIT House_n, 2013] and Homelab [Philips HomeLab, 2013] aim to design home-related products and services for satisfying inhabitants future needs as they live in their homes. In industry, the Microsoft HomeOS [Dixon et al., 2012] provides software interface that aims at shortening the life cycle for developing new applications in order to better control and monitor the homes.

## 2.2 Tracking and Recognizing Activities

Tracking and recognizing individuals in a smart home is a fundamental research area in smart home research. Imagine this scenario: a smart home resident wakes up in the morning, then the bedroom light turns on immediately; he steps into the bathroom, before that the light and shower in the bathroom already turn on for him (the lights turn off when he leaves each area to avoid excessive energy consumption); When he finishes the shower, the display system will prompt him with a message that notifies him to take his medicine.

The technologies to track and recognize individual activities are essential to realize this scenario. To implement various applications, a number of tracking systems have been designed. These systems can be categorized based on the different types of sensors and algorithms that are utilized. Some projects [Lasecki et al., 2013, Chen et al., 2010b, Ryoo and Aggarwal, 2009] identify activities using stereo cameras. The streaming video events provide sufficient detail to track multiple people fast and accurately. However, the drawback of this technique is that the inhabitants may feel uncomfortable when they are directly monitored by the camera. To avoid such a privacy concern, research groups made efforts to track and recognize activities using less intrusive sensor systems, including radio-frequency identification (RFID) tags [Ranjan et al., 2012, Yang et al., 2011, Gu et al., 2009], wearable sensors [Gao

et al., 2012, Pantelopoulos and Bourbakis, 2010, Maurer et al., 2006b], and pressure sensors [Chen et al., 2012, Meyer et al., 2006]. Based on the simple assumption that all the features are conditionally independent, the Naïve Bayes classifier still yields good performance when provided with plenty of sample data. The classifier has been widely applied in activity recognition research [Brdiczka et al., 2007, van Kasteren and Krose, 2007, Bao and Intille, 2004]. The hidden Markov Model (HMM) induces a probability distribution over hidden states that correspond to continuous observed events. In the area of activity recognition, the hidden state can represent the activities the residents are executing and the observed events can be replaced with real-time events captured by the sensors. Thus, many researchers and projects [Guenterberg et al., 2012, Zia Uddin et al., 2010, Tapia et al., 2004] have applied HMMs into their own activity recognition models. Other machine learning models have been considered as well. For example, decision trees have been applied to recognize a logical description of activities by examining acceleration data from wearable devices [Ghasemzadeh and Jafari, 2011, Maurer et al., 2006a]. Furthermore, the CareMedia project [Chen et al., 2005] applies a support vector machine (SVM) to learn and train the combined events from video and audio sensors to identify social interactions for caregiving people.

In the CASAS smart home project [Cook et al., 2009], motion sensors embedded on the ceiling and magnetic sensors placed on doors are used to track and localize the

residents. This dissertation does not introduce a new activity recognition algorithm. Instead, our home energy model associates resident activities with energy usage in the home. The result of our model shows there is a strong relationship between resident activity and energy consumption, which can be used to predict and simulate energy usage in the home setting.

## 2.3 Behavior-based Energy Management

Traditionally, many studies aim at reducing total energy consumption or motivating consumers consumption during off-peak periods with the aid of economic theory. The dynamic pricing of electricity [Joskow and Wolfram, 2012] is one of these methods. In this method the provider changes retail prices dynamically based on the level of time-of-use rates and critical-peak pricing tariffs. However, the study [Ito, 2010] points out that the influence of economic incentives is rather small for regular household. Limited consumer response to economic incentives may be caused by lacking details of energy consumption in the household. Recently, the energy community has come to realize that resident activity is one of main factors that influence the level of household consumption. Faruqui and Sergici [Wilson and Dowlatabadi, 2007] observe that the policy of critical-peak pricing is far more effective when associating pricing with the technologies that allow remote control of electrical appliances when

the consumers receive the details on their own consumption in the house. Moreover, a number of researchers have investigated energy management and control relying on occupancy context information. Roy et. al [Roy et al., 2003] provide a location-aware system to minimize energy consumption in an office environment without sacrificing human comfort. The system optimizes energy consumption by adjusting the usage of electrical devices, such as lights and computers. Their results show that the energy expended can be cut by almost 50% through location-based knowledge. More efforts reported in the literature [Harris and Cahill, 2005, 2007] use Bayesian networks to predict behavior patterns as the context for managing appliances efficiently. However, these methods rely on simple acoustic sensors. Finer-grained sensors may be required to provide sufficient resident information. Another study [Newsham and Birt, 2010] provides an ARIMAX model to predict building energy consumption based on knowledge of the occupants. This model indicates that the level of occupancy is a significant input variable to improve performance. Similarly, Seryak and Kissock [Seryak and Kissock, 2003] associate occupant characteristics, such as occupant behavior, number of residents and how long they stay in the home, with energy consumption in various houses. Using an alternative approach, Reinisch et. al [Reinisch et al., 2011] propose a comprehensive framework that integrates all available types of sensor information into an extensive knowledge base to improve home energy efficiency and user comfort. Through simulating the home setting using their tool, an optimum control strategy

can be achieved to reduce energy expense without reducing human comfort levels.

## 2.4 Non-intrusive Load Monitoring

A customer's power bill provides little detailed information. Instead, consumers see aggregate information such as the total expense and the total energy consumed in the home. In order to provide more detail about the nature of usage, researchers are designing hardware and software solutions to collecting finer-grained usage information. Jiang et. al [Jiang et al., 2009] use a wireless sensor system to monitor energy usage and control devices in a building environment. This approach requires expensive professional installation to measure and monitor each appliance. Alternatively, a non-intrusive appliance load monitor [Hart, 1992] has been designed to detect the turning on and off of individual appliances in an electrical circuit. Instead of installing monitoring sensors, this technology can provide more detailed information of electricity usage in the home environment, such as when and where the energy is being consumed. A few studies have focused specifically on non-intrusive appliance detection. Kato et al. [Kato et al., 2009] extract features from power waveforms using Linear Discriminant Analysis (LDA) and employ support vector machines (SVM) to classify appliances. Gupta et al. [Gupta et al., 2010] analyze frequency electromagnetic interference (EMI) on the power line, and then use SVMs to identify unique

occurrences of switching events. Bauer et al. [Bauer et al., 2009] developed a monitoring sensor, connecting to the power outlet, to analyze which appliances are in use and how the appliances are being used in the kitchen. Researchers [Berges et al., 2011] at Carnegie Mellon University first use a generalized likelihood ratio to detect switch power events, then apply machine learning classifiers to associate these events with their respective appliances. Besides detecting electrical appliances, the non-intrusive load monitor technology has also been extended to other areas, including detection of water activity [Froehlich et al., 2009] and gas usage [Cohn et al., 2010].

## 2.5 Energy Consumption Modeling

A key component of behavior-based energy-efficiency research is to design fine-grained predictive models for individual homes and communities. These models can be used in a number of ways, one of which is to help determine current and future loads to configure power feeders across the power grid. With accurate models to predict use, utility companies can re-distribute the load for the power grid, and can make decisions about new power sources using the models of energy consumption. In the power industry, static load models [Li et al., 2007] have been already widely used to simulate end-use electricity consumption and identify the parameters for transient stability analysis. These models need long-term historical load data and

are only applicable to large urban zones. Some urban climate researchers have used economic survey data to evaluate the level of energy consumption in various building sectors. Sailor and Lu [Sailor and Lu, 2004] associate population density with monthly state-level energy consumption, and further estimate diurnal profiles of anthropogenic heating. Although this method can be applied to other cities, the population density formulation varies greatly and cannot be capable of accurately estimating energy consumption. The main reason is that energy usage relies heavily on many factors, such as diurnal temperature, human behavior, and building characteristics. Narayan and Smyth [Narayan and Smyth, 2005] extend this application by associating more related variables with building energy demand. They found that residential income and electricity price are the most significant factors to determine residential energy consumption.

For individual buildings, a precise model can be helpful to not only evaluate the effect of different energy-efficiency improvement options, but can also encourage occupants to adopt energy-efficient behaviors. For example, the model can identify inefficient and aging heat equipment and household appliances, and suggest that users replace them in order to save electricity cost. Mihalakakou et. al [Mihalakakou et al., 2002] apply neural network models to estimate energy consumption of the building relying on climatic parameters such as air temperature and solar radiation. Using similar neural networks models, Yang et al. [Yang et al., 2005] simulate on-line build-

ing energy consumption as an input of outdoor and water temperature and electric demand. Similarly, Yao and Steemers [Yao and Steemers, 2005] simulate daily energy demand by determining the optimum combination of physical and behavioral factors. Nearly 25% of energy consumption has been consumed by heating, ventilation, and air-conditioning (HVAC) systems in buildings [McQuade, 2009]. To improve the energy-efficient of the HVAC system, Aswani et. al [Aswani et al., 2012b] built two testbeds for modeling and control of the HVAC system. Based on these testbeds, a learning-based model predictive control (LBMPC) technique [Aswani et al., 2012a] is used to reduce energy usage by estimating the heat effects from occupany, solar effects, and outside air temperature. The EnergyPlus model [Pang et al., 2011] uses an Energy Management and Control System (EMCS) to detect building features, weather conditions, and HVAC system use, then simulates whole building energy consumption based on these features. This approach is impractical for large applications due to the prohibitive cost and professional knowledge required for installation. Rather than estimate energy usage of the whole building, some research efforts reported in the literature [Korolija et al., 2013, Griffith and Crawley, 2006] associate building descriptions with energy consumption over a large scale.

In an effort that is similar to ours, Kolter and Ferreira [Kolter and Jr., 2011] combine geographic information system (GIS) and building characteristics to estimate energy consumption for individual buildings. The main difference between this model

and our work is that their models are based upon a data set of monthly electricity and gas bills, but our models are built directly upon real-time smart meter data. In addition, we incorporate more detailed input features, 66 features in total, rather than the 33 features used in this earlier work. In comparison to other past works, this dissertation offers several unique contributions. The real-world data sets used in this work are capable of representing realistic residential patterns. Furthermore, we use machine learning techniques to develop predictive models, and present a public web-based interface for providing feedback contextual information about their energy usage. To the best of our knowledge, this is one of the first works that use a large collection of real-world smart meter data to model residential energy usage.

## 2.6 Anomaly Detection on Energy Data

Anomaly detection finds extensive use in a wide variety of applications such as intrusion detection [Gwadera et al., 2005], fault detection for credit cards [Phua et al., 2004], medical health [Lin et al., 2005], industrial damage detection [Keogh et al., 2007], and sensor networks [Zhang et al., 2007]. Since society is becoming increasingly aware of the impact of energy efficiency, anomaly detection has been directed toward energy aspects of building management.

Statistical methods offer the most popular solutions to detecting anomalies in

energy consumption time series data. Some researchers have applied a variety of statistical methods to analyze energy data in general. Seem [Seem, 2007] extracts features, such as average and peak consumption, from daily energy consumption. Then two statistical models, the generalized extreme studentized deviate (GESD) and Modified z-score [Crosby, 1994], are applied along with an outlier detector in order to identify abnormally high or low energy use. In this dissertation, three classical models ( $\overline{x}$ control chart, boxplot and CUSUM control chart) [Chen and Cook, 2011b] are applied to automatically detect and analyze outliers and trends in household energy usage. However, these methods rely on the assumption that the data is sampled from a particular distribution, which may not hold. It can also be difficult to identify contextual anomalies with small fluctuation.

As a result, machine learning methods are also applied to identify outliers in energy data. Jakkula and Cook [Jakkula and Cook, 2010] apply an unsupervised learning algorithm, K-nearest neighbor (K-NN), to find energy patterns far away from their neighbors. However, this method cannot detect anomalies that have close neighbors but appear rarely. Thus, more work on the topic is needed. Instead of analyzing data in the time domain, Grwtham et al. [Bellala et al., 2011] translate time series data to the frequency spectrum data and then use a K-NN algorithm to identify anomalies in sparse regions of the data. This method can detect abnormal patterns with low frequency, but ignore common anomalies that exhibit particularly

high variation.

In this dissertation, our pattern-based approach [Chen and Cook, 2011a] defines a framework to mine raw energy data by transforming raw energy data into a symbol sequence, and then extends a suffix-tree data structure to create an efficient representation for analyzing structural patterns. Our methods, therefore, are expected to detect three types of anomalies: 1) anomalies that vary dramatically from other patterns; 2) anomalies with large fluctuation; and 3) anomalies that occur rarely. A clustering method is finally applied to detect such anomalies far from their cluster centroids. To the best of our knowledge, this is the first work that applies a pattern-discovering approach into detecting energy outliers in home environments.

In the CASAS project, the ubiquitous and simple sensors are employed to track residents for behavior-based research [Cook et al., 2009]. To better understand the algorithms in this dissertation and their evaluation in the context of a smart home testbed, the design of the CASAS smart homes must first be introduced. This includes explanations of the sensors that are used for tracking individuals, the network that provides sensor communication, and the methods employed for event representation. We provide details on the CASAS smart home implementation in Chapter 3.

# CHAPTER 3. THE CASAS SMART HOME ENVIRONMENTS

With the increasing need to develop technologies that support living at home independently and comfortably, the new technologies in the fields of physical sensors and intelligent algorithms are designed to implement smart homes that are being applied to a wide range of health, family entertainment and sustainability applications. The CASAS Smart Home project designed a common platform that integrates a collection of hardware and software tools in an effort to explore these challenging research goals. In this chapter, the details about the implementation of the CASAS smart home and various types of sensor data collected in this environment will be introduced.

## 3.1 System Architecture

Unlike regular houses, a smart home environment is equipped with sensors which can perceive the surrounding environment and transfers sensor messages through a network to a central computer which reasons about the state of the environment, the state of the residents, and the goal that it is trying to achieve or optimize. The

computer selects an action to take when needed and transmits the action to the hardware or the resident in order to improve their experience in the home setting [Cook et al., 2009]. As shown in Figure 3.1, the system architecture of the CASAS smart home includes three main parts introduced in the following sections.

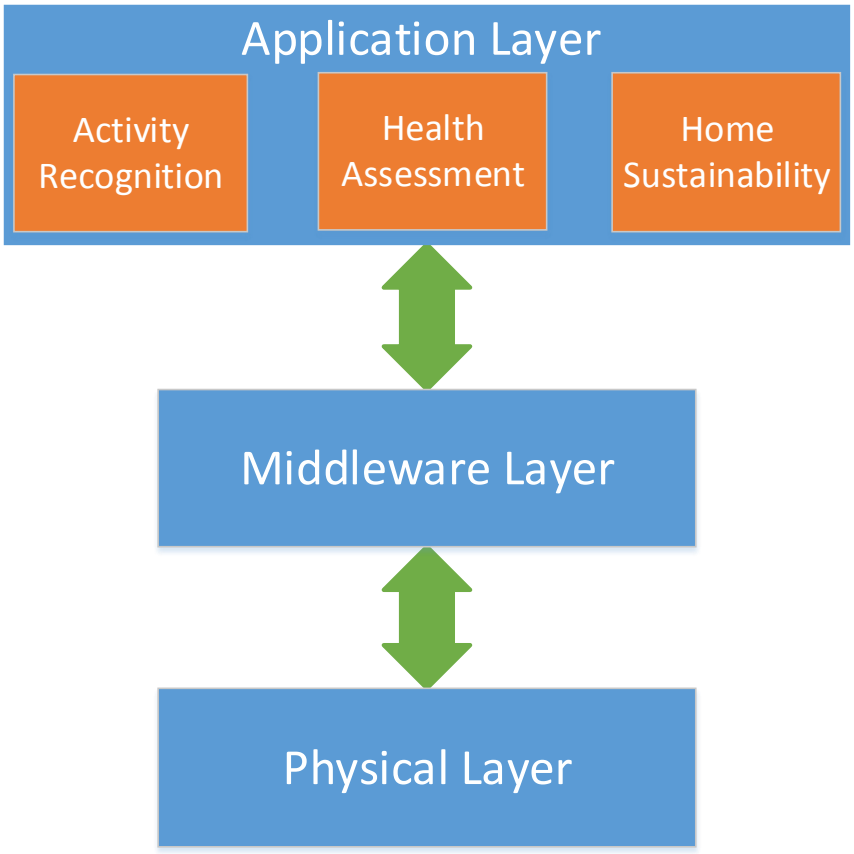



**Figure 3.1:** System architecture of the CASAS smart home project.

### *3.1.1 Physical Layer*

The CASAS physical layer is designed to be robust, easy to install, energy efficient and acceptable. It contains hardware components including sensors and actuators. The whole architecture employs a wired and Zigbee wireless sensor network that communicates directly with the hardware components [Cook, to appear]. The main components of the CASAS physical layers will be introduced as follows.

**Passive Infrared (PIR) Motion Sensor**

PIR motion sensors are installed on the ceiling or the wall in the home for detecting resident movement. The devices attempt to measure heat-based movement from the residents. The passive term means that the device only accepts the infrared light without generating any energy during detection. There are two types of PIR sensors installed in the sensor platform: 1) area sensors and 2) point sensors. Figure 3.2 shows the picture of the CASAS wireless motion sensors. As shown in Figure 3.3, the area motion sensor installed on the wall can be used to detect the residents in a larger range of the space, but provide little information about their exact location.

To get more accurate information, point sensors have a capability of detecting the residents directly below the sensor within a small range. Because of their reliability and accuracy, a large grid of these motions sensors is installed to track human behavior through the whole home. The CASAS smart home installed this type of motion sensor


**Figure 3.2:** PIR motion sensors used in CASAS testbeds.

broadly to cover the whole home for tracking human behavior. Table 3.1 shows sample messages generated by the motion sensors. The **Date** and **Time** columns indicate when the motion event has been generated; the **Sensor ID** shows the identification of the specific motion sensor; the **Message** 'ON/OFF' means the resident is moving or has ceased to move in the space of the motion sensor.

### Magnetic Door Sensors

A magnetic door sensor is a magnetic-based electrical switch. The switch is open when a magnetic field is near; the switch closes when the magnet field is pulled

**Figure 3.3:** PIR motion sensors on the ceiling.

**Table 3.1:** Raw data from motion sensors.

| Date | Time | Senosr ID | Message |
|---|---|---|---|
| 2009-02-06 | 17:17:36 | M45 | ON |
| 2009-02-06 | 17:17:40 | M45 | OFF |

away. The opening and closing of a door, such as home entry doors, room entry doors, cabinets and refrigerators, can be detected using these sensors. The real-time switch status can be sent to the central server automatically. A CASAS door sensor can be placed on the main exterior door of the house as shown in Figure 3.4. Table 3.2

provides a sample message generated by the magnetic door sensors. **Date** and **Time** label the moment the event happens; **Sensor ID** shows the identification of the door sensors; the **Message** OPEN and CLOSE indicate the door has been opened and closed.

**Table 3.2:** Raw data from magnetic door sensors.

| Date | Time | Senosr ID | Message |
|---|---|---|---|
| 2009-02-06 | 19:18:36 | D003 | OPEN |
| 2009-02-06 | 19:18:40 | D003 | CLOSE |


**Figure 3.4:** Magnetic sensor placed on the front door of a house.

**OneMeter Power Metering**

The OneMeter Power Metering device can measure the current energy, wattage, and the total amount of electricity in units of kilowatt hours (KWh), on a single power line. Usually the power metering is installed on the main power line for monitoring energy cost in the whole house. In the smart homes, all the power meter readings can be transferred to the server via the middleware framework automatically. With the help of visualization technology, the residents can get direct real-time feedback about their electric consumption. The format of the CASAS power reading messages is shown in Table 3.3.

**Table 3.3:** Raw data from power metering.

| Date | Time | Senosr ID | Message |
|---|---|---|---|
| 2009-02-06 | 11:18:37 | P001 | 930W |
| 2009-02-06 | 11:20:40 | P001 | 0.4 KWh |

## 3.2 Middleware Layer

In the CASAS architecture, the physical layer provides various types of sensors to monitor residents behaviors and condition in the homes. All these sensors are connected by the middleware component for transmitting the newest events to the

server in real time. The middleware layer is governed by a publish/subscribe manager. This central manager provides signal channels that allow other sensor components to publish and received these messages, and then forwards them to the agents. In addition, this middleware also provides extra services, such as associating real-time stamps to events, assigning unique identifiers, monitoring and maintaining sensor states. More details of the middleware layer are documented in the literature [Kusznir, 2010].

## 3.3 Application Layer

Based on the sensor framework and middleware layer, the CASAS project developed various applications to provide capabilities with no customization or training. Four core software components and applications have been designed to meet this goal.

**Activity Recognition**

Sensor events include a date, a time, and a message corresponding to a sensor reading. The goal of activity recognition is to associate a sequence of sensor events with corresponding activity labels. To perform this task, a hidden Markov model [Singla et al., 2009] has been implemented to track various activities in complex settings. The model can provide the possible activities which are being executed consistent with the current sensor events. The result of the recognition rate can be

as high as 95%. To track multiple residents' activities, Crandall and Cook [Crandall and Cook, 2011] extend the hidden Markov model to recognize activities of multiple residents by identifying their own unique patterns. However, both of these approaches require manually-provided activity labels for training. How to handle unlabeled data is a new challenge for recognizing activities. To solve this challenge, Rashidi et. al [Rashidi et al., 2011] design an unsupervised learning algorithm to discover activities from unlabelled sensor events.

**Health Assistance**

The population of the world is aging. By 2050, the number of people over 85 is expected to triple [Vincent and Velkoff, 2010]. Innovative and preventive health assistance methods need to be developed for elderly people within their own home. The CASAS smart homes have the capability of monitoring changes in behavior over multiple years. Specific physical health parameters are monitored, including activity level, sleep quality, and time spent in various rooms. A web-based visualization of these resident parameters is shown in Figure 3.5.

In addition, the PUCK prompting system [Das et al., 2012a] is an activity-aware health assistive tool that prompts individuals to initiate important daily activities such as taking medicine, exercising, or calling a family member. Considering the portability of a mobile platform, the Android-based mobile solution [Das et al., 2012b]


**Figure 3.5:** Activity monitoring for smart home residents.

is designed to provide real-time prompting within smart environments. To detect and prevent potential illness, health assessment research is being conducted to detect changes in health based on sensing changes in behavioral patterns.

### Home Sustainability

The potential relationship between human activities and energy usage in the home environment has been inspected. Classification and regression models [Chen et al., 2010a, Chen and Cook, 2012] have been designed to explore this relationship

and estimate the level of energy consumption corresponding to current activities. Anomalies in energy data also need to be identified in order to detect abnormal individual power readings and unusual power patterns hidden in the regular smart home sensor events. To detect individual anomalies, several classical statistic models [Chen and Cook, 2011b] are applied to examine time-series sensor events. However, these statistic models have no capability of capturing the context anomaly that is composed by several consecutive data points. To address this challenge, pattern discovery and clustering methods [Chen and Cook, 2011a] have been explored. These works are all included in this dissertation.

### *3.3.1 Kyoto and Tulum Smart Environments Description*

In this dissertation two CASAS smart environments (Kyoto and Tulum) are used to collect sensor events to analyze energy consumption in smart home environments. As shown in Figure 3.6, the Kyoto smart home apartment testbed consists of three bedrooms, one bathroom, a kitchen, and a living/dining room. The two-floor Tulum apartment is also instrumented, which consists of two bedrooms, a living room, dining room, kitchen, and a bathroom. The circles in the figure represent the positions of the motion sensors. The motions sensors facilitate tracking the residents who are moving through the space. In addition, the testbed also includes temperature

sensors as well as custom-built analog sensors to provide temperature readings and hot water, cold water and stove burner use. A power meter records the amount of instantaneous power usage and the total amount of power which is used. An in-house sensor network captures all sensor events and stores them in a SQL database for long-term storage.

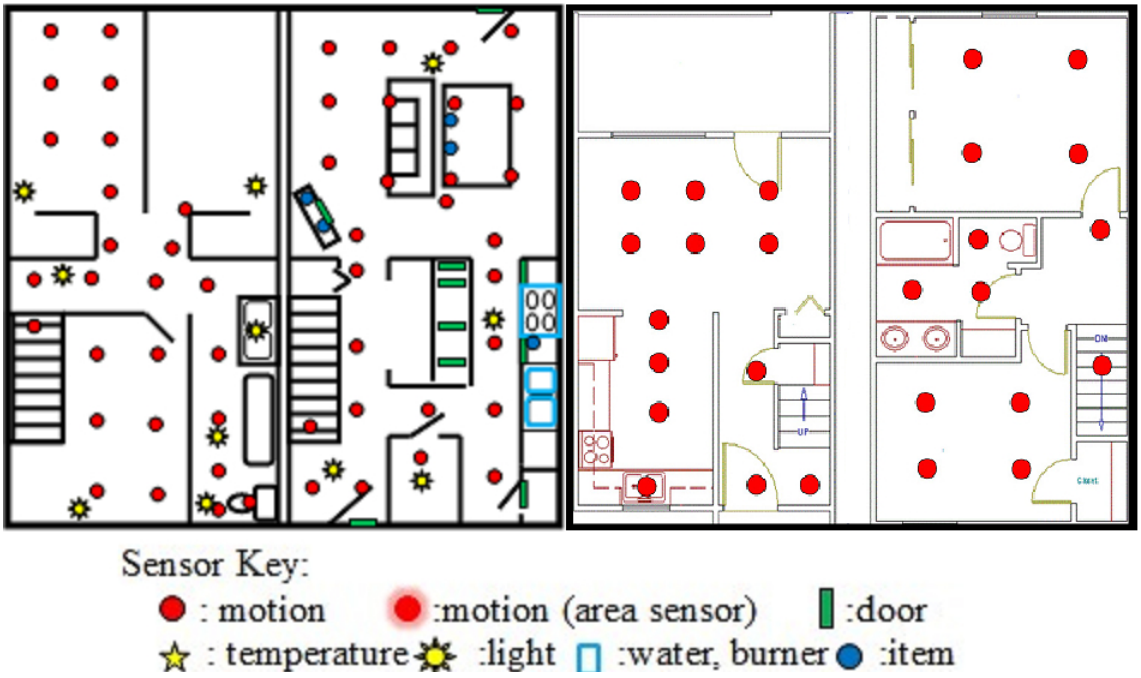



**Figure 3.6:** CASAS smart apartment testbeds: Kyoto (left) and Tulum (right).

Our research team has installed 38 of these types of smart homes around the Pacific Northwest at a cost of approximately $3,000 per home. Installation takes 2-3 hours and removing the smart home equipment takes about 30 minutes. The components are not intrusive and the residents often forget they are present after the

first week. As a result, using this technology is a fairly realistic approach to provide context-aware services and investigating the link between behavior and sustainability.

## 3.4 CASASviz - Web-Based Smart Home Visualization

Smart homes are equipped with a variety of sensors to collect data about residential behaviours. All of these sensors generate huge volumes of data when they detect residents performing activities. Such voluminous raw data provides us with no insights until they are analyzed systematically. Systematic analysis may refer to extraction of information by mining data online or offline and visualizing information for easier interpretation.

Here we introduce the first contribution we made to the analysis of energy consumption using smart home technologies. Specifically, we provide details of our CASASviz visualization algorithm. The CASASviz tool [Chen and Dawadi, 2011] is originally designed to improve the accessibility of smart environment technology to the caregivers by creating a user-friendly interface to represent information gathered from the smart home. In the context of this dissertation, by integrating learning techniques, we design CASASviz to be a tool for consumers and smart home residents, by making sensor data easier to understand, and in turn, to better assist in guiding users in energy-efficient behaviors. Because it offers a web-based solution, the CASASviz

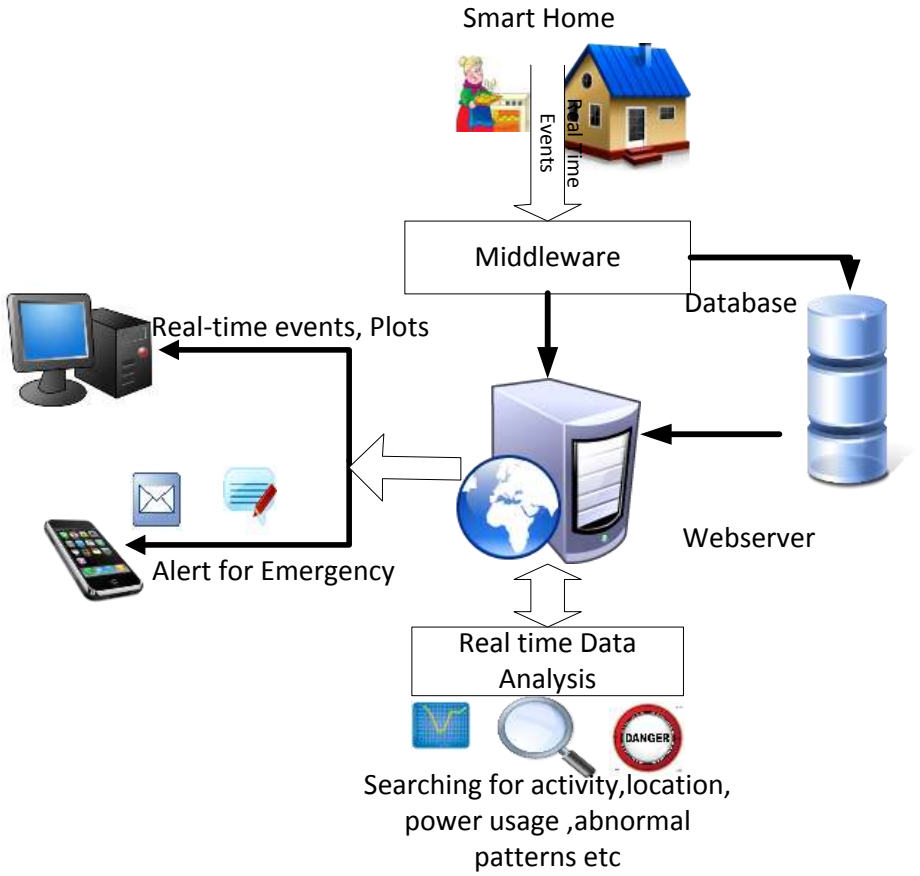



**Figure 3.7:** Depiction of the CASASviz system architecture.

system can be accessed from a computer or from a smart phone, which allows the user to monitor their homes and receive alerts remotely.

### *3.4.1 System Architecture*

Figure 3.7 illustrates the system architecture of CASASiviz for resident use. The middleware collects and records the sensor events from different types of sensors in the smart home. All the sensor events are stored in a SQL database for future analysis. The middleware also sends real time events directly to the webserver. As an option, our system also supports visualizing data that is stored in the database. The webserver module is responsible for communicating with the middleware and SQL database and transmitting the sensor events to the client side for generating the user interface. The data analysis module uses several machine learning and data mining techniques to learn resident patterns and activities being conducted in the smart home. It also detects abnormal behaviors and frequently occurring patterns of the resident, and builds various types of plots that show the residential behavior in the smart home. The residents can monitor their houses using computers or smart phones. They can also receive messages by email or SMS text message when the system detects emergency situations.


**Figure 3.8:** Main CASASviz visualizer interface.

### *3.4.2 Tracking the daily activities in real time*

CASASviz is designed so that occupants can get real time information about their homes and also important information about the residents' current status such as location, temperature, and the activity being performed. As shown in Figure 3.8, circles show the location of the motion sensors, and a red circle denotes the resident is triggering a motion sensor when he walks through this sensor. For real time data communication, the system uses a push technique called Comet [Crane and McCarthy, 2008]. Unlike traditional request/response technologies such as Ajax [Powell, 2008], in which the client wastes time waiting for a response from the web

server, Comet pushes the data automatically to a client browser from a previously opened connection. Thus, the client does not need to poll the server to fetch new data, which greatly improves the speed of the visualization. Finally, our system implements online activity recognition, location recognition, and activity level monitoring during a specific period of time.

### *3.4.3 Sensor Heat Map*

CASASviz can display a heat map that shows the amount of activity of the resident in areas of the home over a period of time, as depicted in Figure 3.9. Colors of the motion sensor icons depict the area of the house where residents spent more time, while the colors of door and item sensors show frequency of usage. The information such as in which rooms residents were more focused, or what type of item they frequently use or do not use, can easily be inferred.

### *3.4.4 Power Usage Visualizer*

In smart environments, power usage is also an important features that reflects behavioral patterns of the residents. As shown in Figure 3.10, CASASviz provides an energy usage visualizer to express energy fluctuations that occurred during defined


**Figure 3.9:** Mobility Heat Map in CASASviz.

user-specified time period. This graph can be used to identify trends and anomalies of power consumption. The power usage visualizer runs in synchrony with the smart home activity visualizer, so that residents can identify behaviors that cause increases or decreases in electricity consumption.

### *3.4.5 Prompting alerts and timely feedback*

Users want to be notified when any emergency situation occurs in the home. In particular, using anomaly detection algorithms, customers can be notified that they

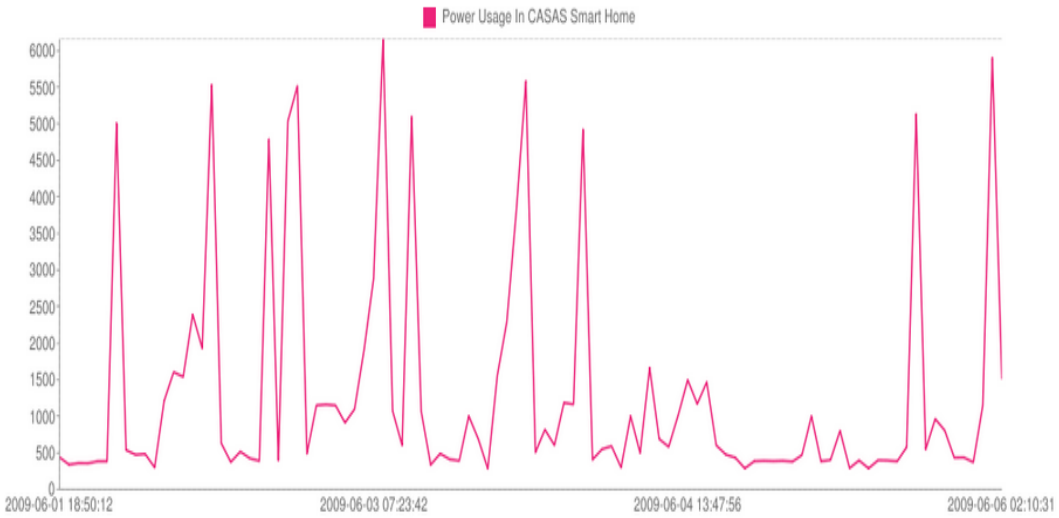



**Figure 3.10:** Power Usage Visualizer in the CASASviz.

consume an exceptionally large amount of energy in a timely manner. More detailed information can also be provided, including when customers performed certain activities, which rooms they occupied, and what appliances they used most frequently during that period. This information can be transmitted to customers in a timely fashion via phone, email or the Internet. Considering these requirements, we provided an ability to detect such situations and send a short message to their mobile device or a message to their email account.

### *3.4.6 Working on Smart Phones*

Smart phones are different from the PC due to the particular screen size and resolution on which the interface should be re-designed for good fitting. Considering this distinct difference, we designed a new user interface focusing on the requirement of smart phones.


(a) Track Residents (b) Monitor Energy

**Figure 3.11:** Mobiledevice-based CASASviz.

Figure 3.11 shows the CASASviz interface on an iPhone device. The purpose of this interface is to display all the information on a smart phone screen, while providing a friendly user experience for the caregiver, especially older caregivers. We configure

the user interface of the system so that it would work for both smart phones and personal computers automatically.

# CHAPTER 4. BEHAVIOUR-BASED ENERGY PREDICTION

We hypothesize that energy consumption in the home is correlated highly with the type of residential activities that are performed and can therefore be predicted based on the activities that occur in smart environments. In this chapter, we validate our hypothesis by introducing two types of prediction methods that predict energy usage based on sensor events collected by a smart home. In the first method, energy data is discretized into several equal-sized classes and the classification methods are used to predict these classes given information about that activity that a resident performs. In the second method, regression models are used to predict numeric energy data directly as a function of various sensor events without knowing the activities the residents are performing.

## 4.1 Classification Energy Prediction

Activity recognition techniques are prevalent in the literature [Kim et al., 2010] and are becoming more robust. Activity recognition algorithms offer a practical approach to automatically correlating activities in the home with energy consumption.

We define the features listed in Table 4.1 to describe an activity performed by an inhabitant in a smart home.

The input to the learning algorithm is a vector of values for the features listed in Table 4.1 as computed for a particular activity that was performed. The output of the learning algorithm is the amount of electricity that is predicted to be consumed while performing the activity. In this chapter, by applying equal-width binning, the target average energy data was discretized into several interval sizes (two classes, three classes, four classes, five classes, six classes, and seven classes) and these classes are treated as the target variable for a supervised learning algorithm.

To train the machine learning algorithms, sensor events were annotated with ground truth consisting of the corresponding activities being performed while the sensor events were generated. All of the activities that the participants perform have some relationship with measurable features such as the time of day, the participants movement patterns throughout the space, and the on/off status of various electrical appliances. These activities are either directly or indirectly associated with a number of electrical appliances and thus have a unique pattern of power consumption. Table 4.2 lists the appliances that are associated with each activity. It should be noted that there are some appliances which are always in use at varying levels, such as the heater (in winter), refrigerator, phone charger, etc. Thus, we postulate that the activities will have a measurable relationship with the energy usage of these appliances as well.

**Table 4.1:** Data features input to classification models.

| **Feature Name** | **Description** |
|---|---|
| *Activity label* | the types of the activities the residents perform |
| *Activity length (seconds)* | the time length of the activity |
| *Day of week* | the current day of week (Monday, Tuesday, Wednesday, Thursday, Friday, Saturday, and Sunday) |
| *Weekday/Weekend* | a binary variable to determine whether the current day is a weekday or weekend |
| *Time of day* | different time slots (morning, noon, afternoon, evening, night, and late night) |
| *Times of individual sensors triggered* | times of different motion sensors that were activated during the activity |
| *Number of motion sensors activated in various rooms* | the number of motion sensors that were triggered in various rooms |
| *Total number of motion sensor events triggered* | the total number of motion sensor events that were triggered during the activity |

**Table 4.2:** Electrical appliances associated with each activity.

| **Activity** | **Appliances Directly Associated** | **Associated Appliances** |
|---|---|---|
| Work at computer | Computer, printer | Localized lights |
| Sleep | None | None |
| Cook | Microwave, oven, stove | Kitchen lights |
| Watch TV | TV, DVD player | Localized lights |
| Shower | Water heater | Localized lights |
| Eating | TV | Localized lights |

### *4.1.1 Analysis of Resident Activities and Energy Usage*

Before we introduce the machine learning algorithms, we first visualize power consumption data for our testbeds that corresponds to the activities being tracked. Figure 4.1 shows the energy fluctuation that occurred during a single day on June 2nd, 2009 in the Kyoto testbed. Resident activities are indicated by the arrows in the figure. The length of the arrows indicates the duration of time for each activity. Note that there are a number of peaks in the graph even though these peaks do not always directly correspond to a known activity. These peaks are due to the water heater, which has the highest energy consumption among all of the appliances in

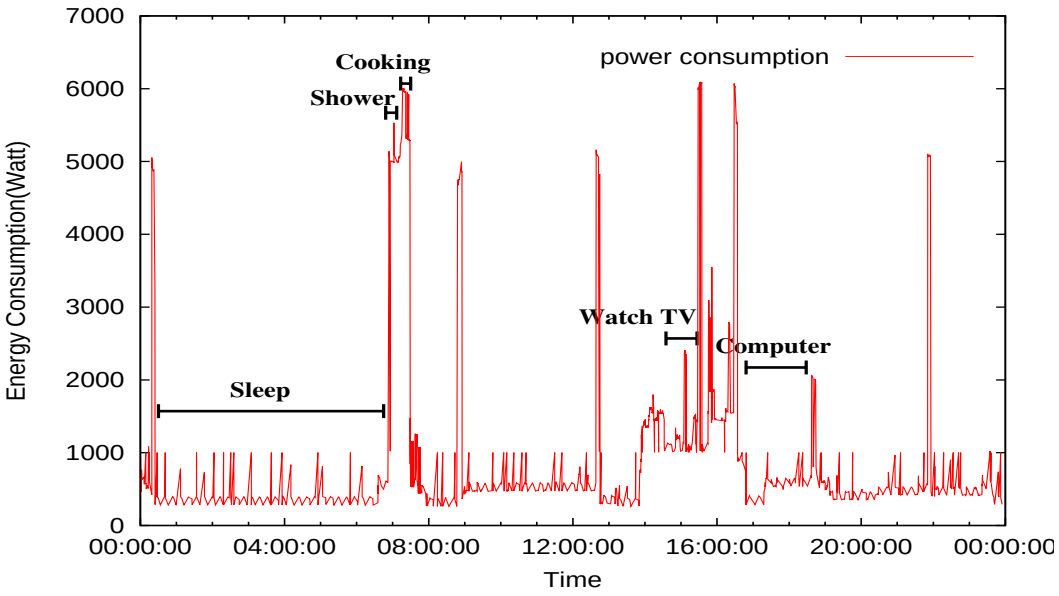



**Figure 4.1:** Energy usage for a single day.

the apartment, even though it is not controlled directly by the resident. The water heater starts heating by itself whenever the temperature of water falls below a certain threshold and is indirectly affected by resident activities that use hot water such as a shower or washing dishes.

Figure 4.2 plots typical energy data for each activity together with the result of applying curve fitting to the data. Curve fitting [Coope, 1993] is the process of building a mathematical function model that can best fit to a series of data points. It serves as an aid for data visualization, and to express the relationships between different data points. From the figure, we see that each activity generates a different

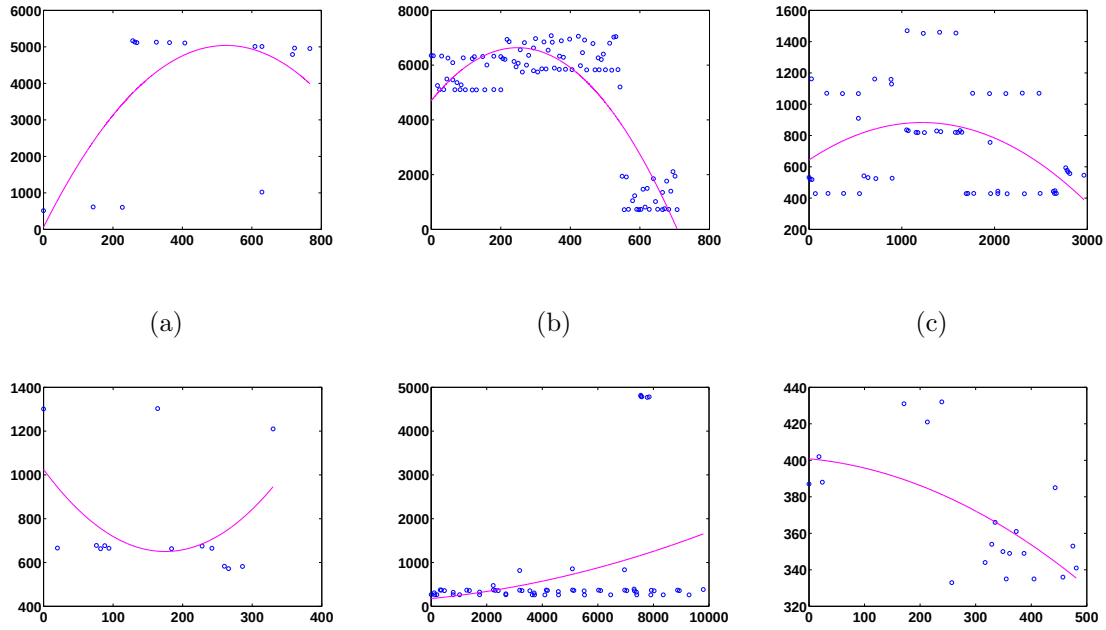


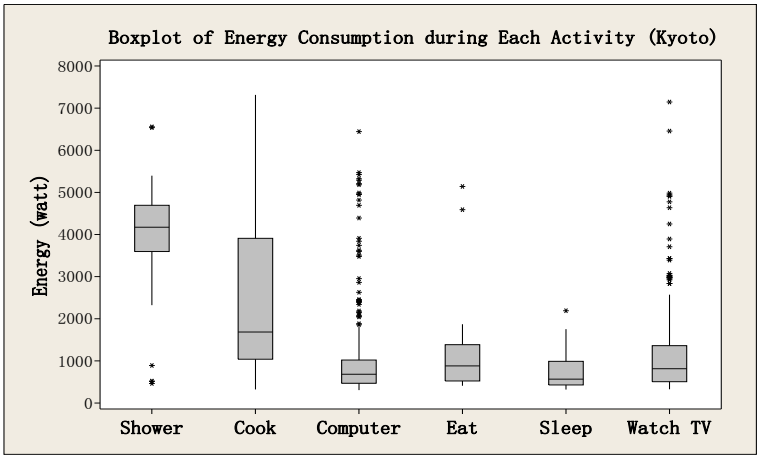

**Figure 4.2:** Energy data curve fitting for each activity. There is a separate graph for each activity: a=shower, b=cook, c=work on computer, d=eat, e=sleep, and f=watch TV. The x-axis in the graphs represents wattage and the y-axis represents time of the activity in seconds.

energy consumption pattern.

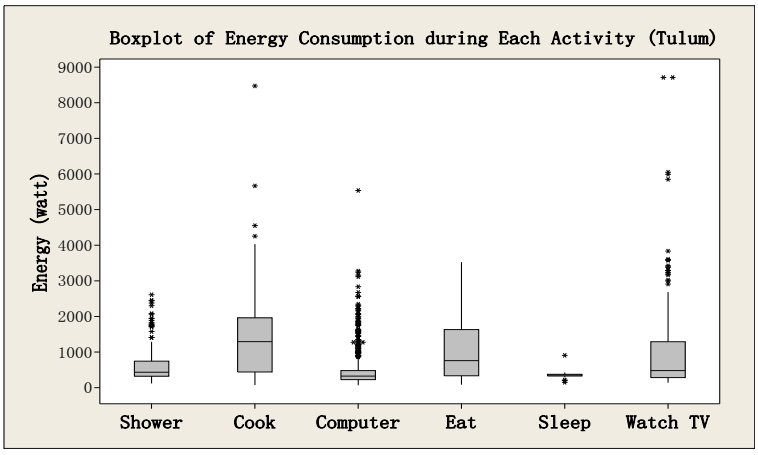
Figure 4.3 illustrates two boxplot graphs of energy consumption for each activity in the Kyoto and Tulum testbeds. Again, the graphs show that each activity utilizes very different amounts of power. In Kyoto, the shower activity consumes the highest amount of energy because the water heater is a larger power consumer. However, the cook activity uses the most energy in Tulum. Cooking in Tulum involves frequent access to the refrigerator and stove, which increases power consumption. Meantime, in both testbeds when the participants were sleeping the energy consumption was the lowest because most appliances were idle.

To compare the difference between residental behaviors on highest and lowest energy hours, the two-month Kyoto dataset is separated into small independent pieces by hour. Each hourly dataset includes motion sensor events and according energy usage (kWh). It should be noted that the residents might go outside and sleep at night. During these periods, no any event can be recorded by the sensors. Thus, such kinds of no-activity datasets were removed. Then, the remaining datasets were sorted based on energy consumption. For the simplicity, the top and bottom 100 hourly datasets, occupying approximately 10% of the whole datasets, were selected for the comparison.

Figure 4.4 denotes the proportion of the total sensor events occurred in each room during the highest and lowest energy hours. Figure 4.4(b) denotes the residents

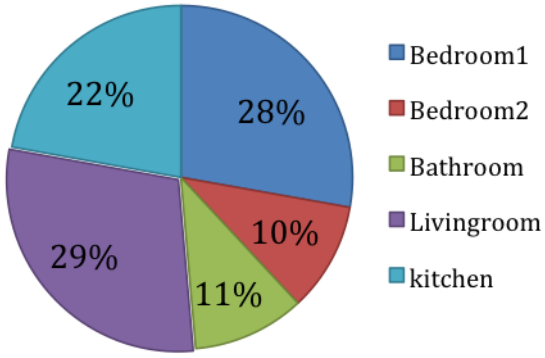



(a) Kyoto

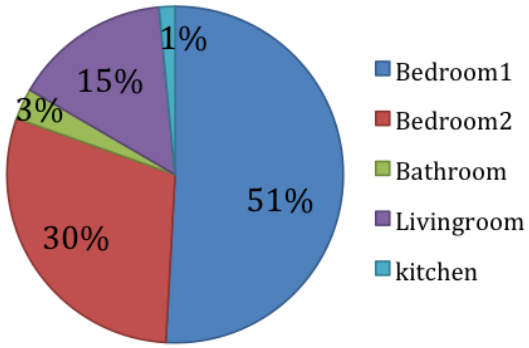



(b) Tulum

**Figure 4.3:** Boxplot of energy data generated by human activates in the Kyoto and Tulum testbeds.




(a) Highest Energy Hours




(b) Lowest Energy Hours

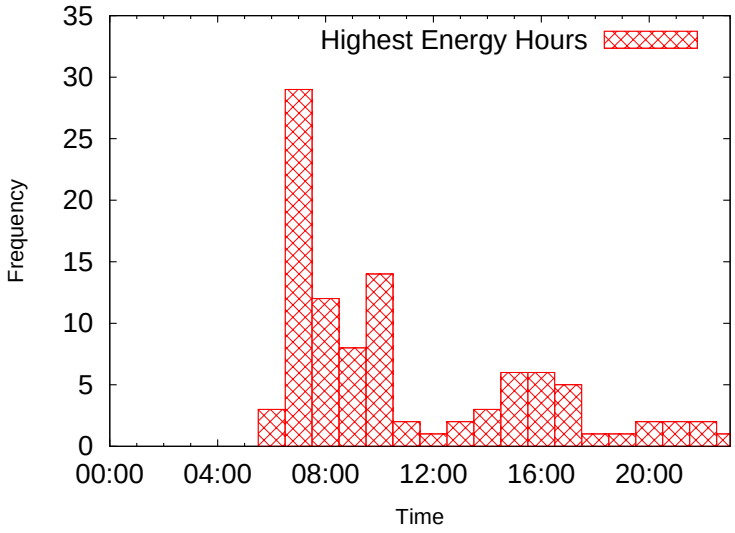
**Figure 4.4:** Distribution of sensor events on various rooms.

spent more than 80% time living in the Bedrooms during the lowest energy hours. During the highest energy hours, only 38% activities occurred in the bedrooms as shown in Figure 4.4(a). The proportion of the kitchen rose from 1% to 22%, and the proportion of the bathroom also had changed to 3% from 11%. It is reasonable

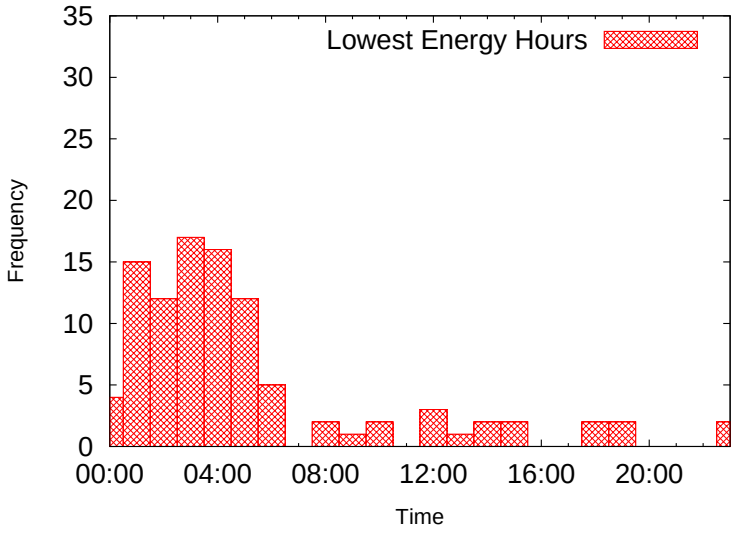
because many large appliances locate in the kitchen such as microwave, oven, and toaste. Besides, water heater can consume particularly larger energy consumption while they have a shower in the bathroom.

Figure 4.5 shows sensor event histograms on various hours. As shown in Figure 4.5(a), the times with highest energy consumption often occur during 06:00 am to 10:00 am. During that time, the residents may get up and cook breakfast. A plenty of the appliances may be used directly and indirectly such as the lights, microwave, and water heater. Figure 4.5(b) shows the residents spent lowest energy consumption during 01:00 am to 06:00 am. During that time, the residents likely go to sleep in the bedrooms.

To help assist in understanding how behavioral patterns influence on energy usage, a web-based pattern visualizer is developed to address this issue. Figure 4.6 denotes the top five behavioral patterns during highest energy hours. Four of these patterns located at the kitchen. There are several large appliances around these patterns such as microwave and oven. These appliances can be identified to be responsible for large energy consumption. To reduce energy usage, the residents can be notified that they can replace these old appliances with more energy-efficient ones. It is surprised that one pattern happening in the bedroom. It can be seen that the resident living in this bedroom may use some unknown appliances, which cost much power. The top five behavioral patterns during lowest energy hours are visualized as




(a) Highest Energy Hours




(b) Lowest Energy Hours

**Figure 4.5:** Distribution of sensor events on various hours.


**Figure 4.6:** The top five behavioral patterns during highest energy hours.


**Figure 4.7:** The top five behavioral patterns during lowest energy hours.

shown in Figure 4.7. All of them are located in the bedrooms. The residents have a high probability of lying in the bedsand most of the appliances are idle.

### *4.1.2 Modeling of Activity-Based Energy Usage*

In the previous section we highlighted the variation in electricity consumption that occurred for different testbeds and activities. In this section we introduce machine learning algorithms that can be used to predict consumption from information about resident activities. Machine learning algorithms are capable of learning and recognizing complex patterns based on sensor data. In this work, the strengths of alternative machine learning algorithms are compared to map these activity features onto a class label indicating the amount of energy that is consumed in the smart environment while the activity is performed. Three popular machine-learning methods were leveraged for this work: a Bayes belief networks classifier, a support vector machine, and a neural network.

Bayesian belief networks (BBNs) [Pearl, 1988] represent a set of conditional independence assumptions by a directed acyclic graph, whose nodes represent random variables and edges represent direct dependence among the variables and are drawn by arrows labeled with the variable name. Unlike the naive Bayes classifier, which assumes that the values of all the attributes are conditionally independent given the

target value, Bayesian belief networks apply conditional independence assumptions only to a subset of the variables. They can be suitable for small and incomplete data sets and they incorporate knowledge from different sources. After the model is built, they can also provide fast responses to queries.

Support Vector Machines (SVMs) [Boser et al., 1992] are a class of training algorithms for data classification, which maximize the margin between the training examples and the class boundary. A SVM learns a hyper-plane which separates instances from multiple energy usage classes with maximum margin.

Artificial Neural Networks (ANNs) [Zornetzer, 1995] are abstract computational models based on the organizational structure of the human brain. The most common learning method for ANNs, called Backpropagation, performs a gradient descent within the solution vector space to attempt to minimize the squared error between the network output values and the target values for these outputs. Although there is no guarantee that an ANN will find the global minimum and the learning procedure may be quite slow, ANNs can be applied to problems where the relationships are dynamic or non-linear and capture many kinds of relationships that may be difficult to model by other machine learning methods. In our experiment, we leverage the Multilayer-Perceptron algorithm with Backpropagation to predict electricity usage.

One challenge that we face when learning a mapping from activity features to energy usage is that the class distribution is highly skewed, as is shown in Figure

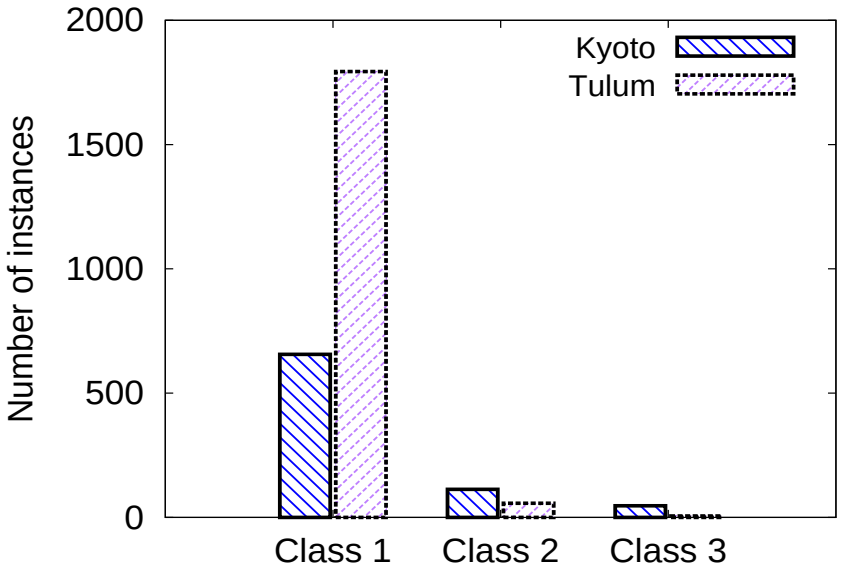



**Figure 4.8:** Distribution of instances in the three energy classes for Kyoto and Tulum.

4.8. This is because most home-based activities require a moderate amount of energy usage, while a small number of activities require substantially more power. Machine learning algorithms may be confounded by this skewing because models that map all (or most) of the cases to the low-energy values will achieve high accuracy, but clearly will not learn the true mapping of activity features to energy usage.

To deal with this imbalanced data, we incorporate a data sampling technique, called SMOTE (Synthetic Minority Over-sampling Technique) [Chawla et al., 2002]. We combine under-sampling methods (to reduce data points in over-represented classes) with over-sampling methods (synthetically generating points for under-represented

classes) to address this problem by using a combination of both under and over sampling, but without data replication. Here over-sampling is performed by synthesizing a new sample corresponding to each minority class by randomly choosing from the point's nearest neighbours. Generation of the synthetic sample is accomplished by first computing the difference between the feature vector (sample) under consideration and its nearest neighbour. Next, this difference is multiplied by a random number between 0 and 1. Finally, the product is added to the feature vector under consideration. The result is a new sample similar to, but not a replica of, the existing data. Under-sampling is performed by randomly removing samples from the majority class until the size of the minority class is close to the size of the majority class. By combining under-sampling and over-sampling, the bias of the machine learning model will lean more toward the minority class.

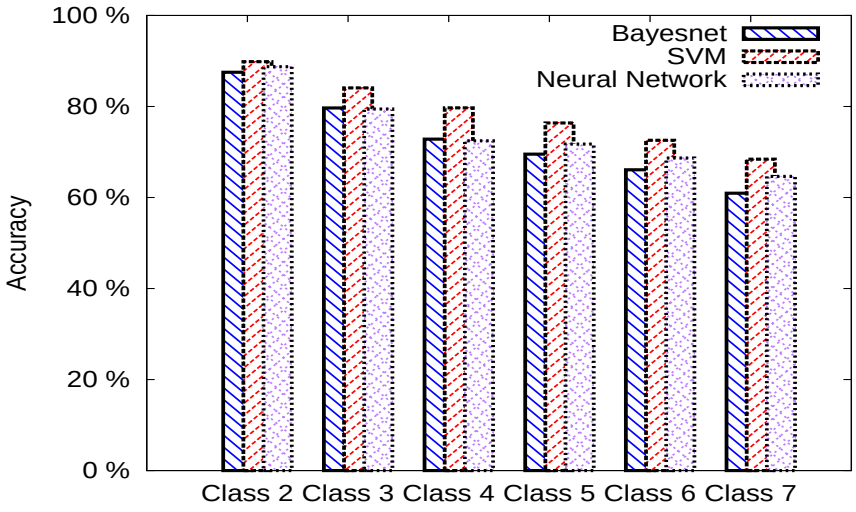
### *4.1.3 Experiment Result*

Two series of energy prediction experiments were performed. The first experiment uses sensor data collected during two months in the Kyoto testbed. In the second experiment, we collected data of two months in the Tulum testbed. Using the Weka machine learning toolset [Witten and Frank, 2005], we assessed the classification accuracy of our three selected machine learning algorithms and reported the

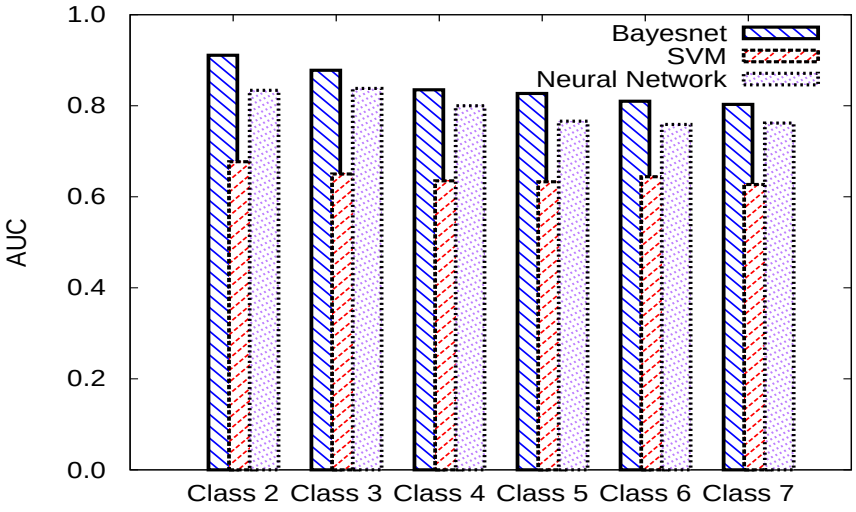
predictive accuracy results based on a 3-fold cross validation. It should be noted that the instances of the class in test group follow the real distribution to examine the performance of the sampling technique.

Conventional performance measures consider different classification errors as equally important. However, this assumption is not practical for our energy prediction, where the class distribution is highly skewed. Therefore, we consider two metrics that measure different aspects of performance. The first metric we use evaluates the conventional accuracy of the classifiers and the second measurement is the area under a ROC curve (AUC), which evaluates overall classifier performance without taking into account class distribution or error cost.

Figures 4.9 and 4.10 plot the accuracies and ROC areas for two different group experiments. The accuracy peaks around 90% for both datasets when predicting the two-class energy usage and the lowest accuracy is around 70% for the seven-class case in both datasets. These results also show that higher accuracy will be attained with a lower precision. Increasing the precision of classification by adding labels, the accuracy across all three algorithms decreases predictably. From the figures, we see that the Bayes network performs worse than the other two classifiers. This is because it is based on the simplified assumption that the features that we use are not conditionally independent. For example, the motion sensors associated with an activity are used to find the total number of motion sensor events is triggered and

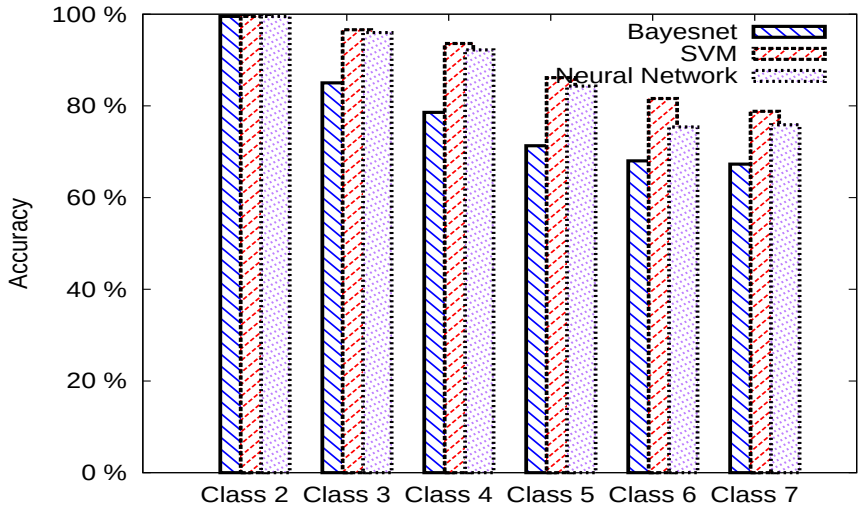



(a) Accuracy

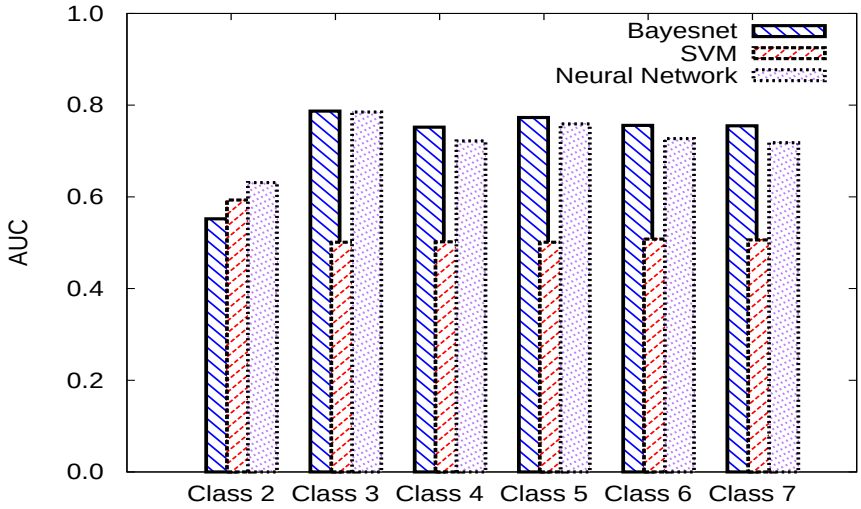



(b) AUC

**Figure 4.9:** Comparison of the accuracy and AUC for the Kyoto dataset.




(a) Accuracy




(b) AUC

**Figure 4.10:** Comparison of the accuracy and AUC for the Tulum dataset.

also the kinds of motion sensors involved in the activity.

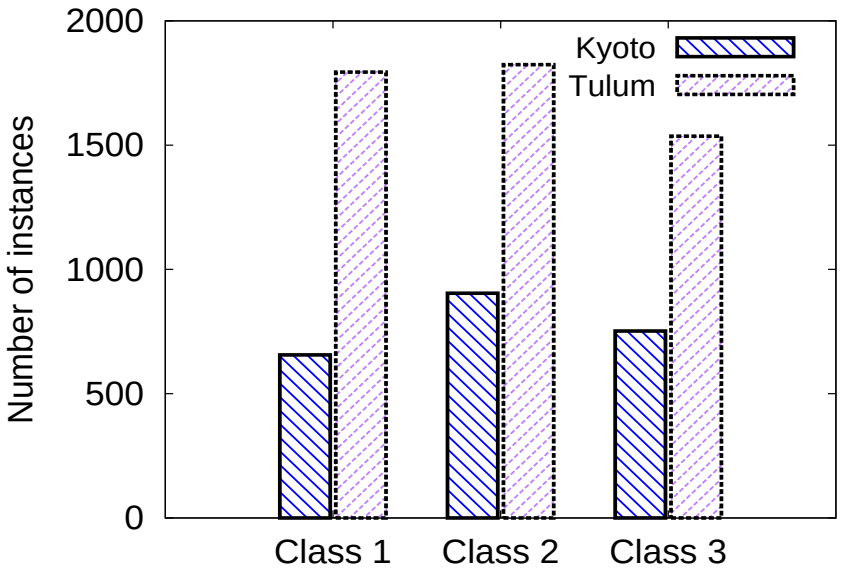
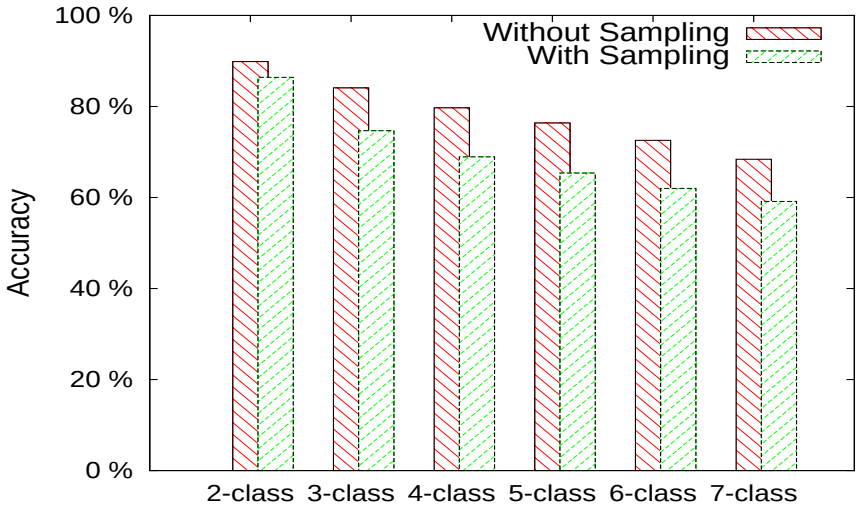
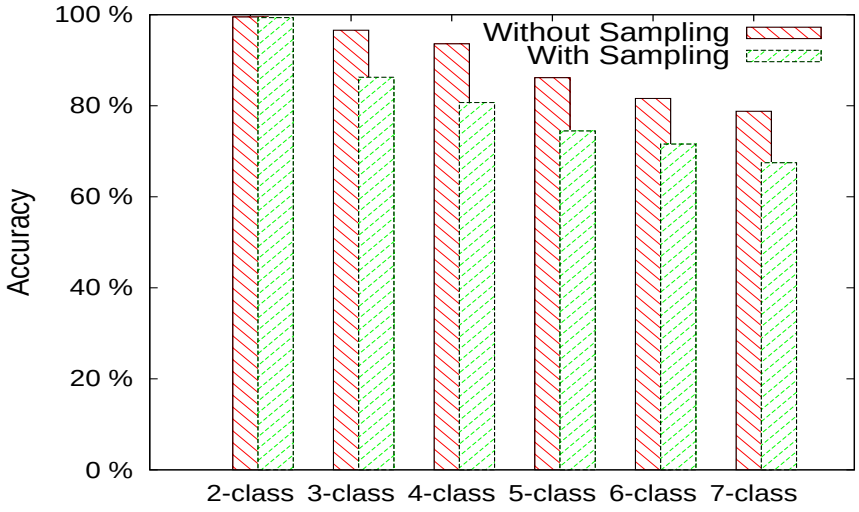



**Figure 4.11:** New distribution of instances in the three energy classes for Kyoto and Tulum.

The ROC curve is not as strong of a measure as accuracy for this experiment. That is because the overall performance is difficult to measure when the training dataset is skewed highly. To deal with these imbalanced datasets, we apply the SMOTE sampling to rebalance our datasets by increasing the number of minority class instances, thereby enabling the classifiers to learn more relevant rules for the minority class. Figure 4.11 depicts the new class distribution of the Kyoto and Tulum datasets after applying SMOTE. Comparing Figures 4.8 and 4.11, we see that two other minority classes have been increased greatly after balancing the datasets.

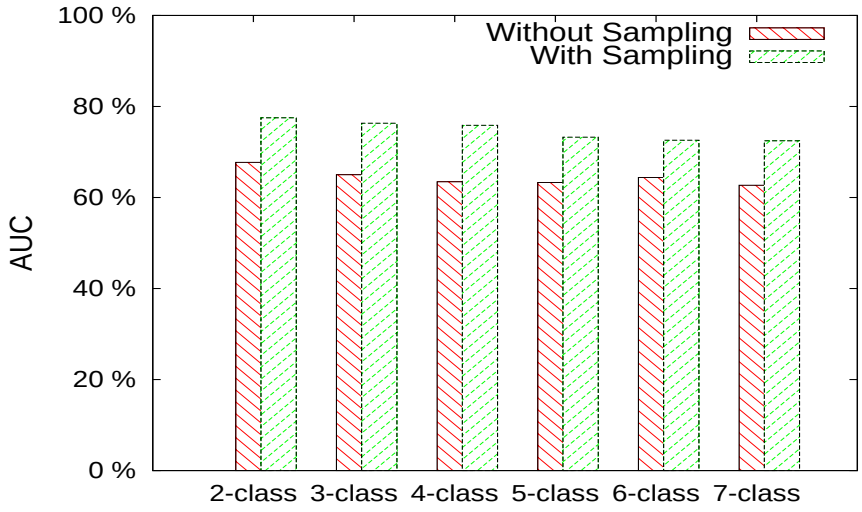



(a) Kyoto

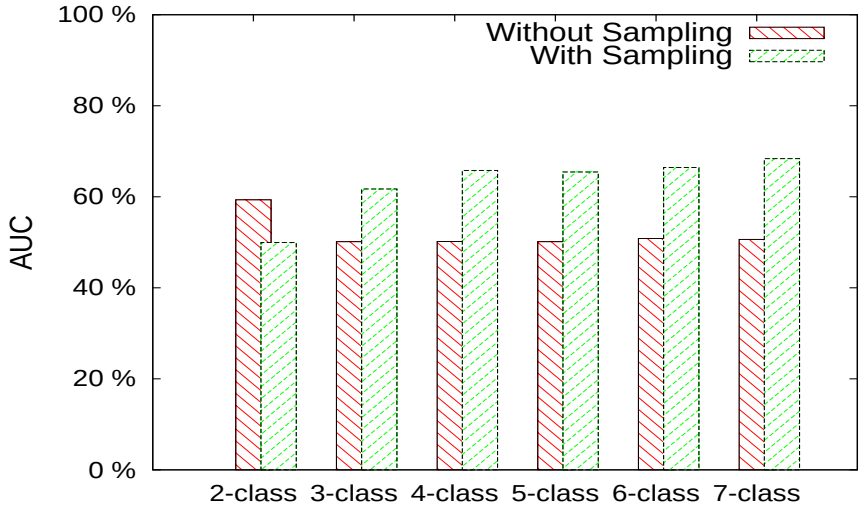



(b) Tulum

**Figure 4.12:** Comparison of the accuracy with and without sampling.




(a) Kyoto




(b) Tulum

**Figure 4.13:** Comparison of AUC with and without sampling.

To analyze the effectiveness of the sampling technique, we evaluate the accuracy of a SVM prediction algorithm on both datasets with and without the sampling. The results are shown in Figures 4.12 and 4.13. In Figure 4.12, the accuracy for both datasets decreased slightly. On the contrary, Figure 4.13 depicts the performance as measured by the area under the ROC curve, which has been improved. After sampling, the classifiers improved the performance to classify the minority class with the loss of decreasing the accuracy. The experimental results show that the sampling technique is a good approach to rebalance the energy data and further improve prediction performance on minority class.

Figures 4.12 and 4.13 also compare the performance of the support vector machine for two different environments, Kyoto and Tulum. Looking at the graphs, Tulum yields a slightly improved performance over Kyoto. This is likely due to the fact that some energy-intensive devices such as room heaters were used in Kyoto but not Tulum (heat is handled from a separate building source in Tulum). These devices are not under the direct control of residents, nor are they directly impacted by activities.

## 4.2 Regression Energy Prediction

In the previous section, the classification algorithms were used to predict energy usage based on the activities of the residents. There are still several problems need to be addressed: First of all, the activities are still difficult to be recognized automatically with high accuracy. In addition, the users may be more interesting in energy consumption during some specific time period. Lastly, numerical values may represent the level of energy consumption more clearly. Based on these considerations, regression models were used to predict energy usage based on sensor features given various time windows.

### *4.2.1 Data Features*

Before illustrating the predictive techniques we use for our data sets, we summarize the specific features we extract from raw data as well as the feature selection techniques we use to identify the maximum relevance features for this paper. The features used to predict energy are shown in Table 4.3.

To predict energy consumption, we assume that energy consumption has a measurable relationship with residents routine activities, which can be detected using motion sensors installed on the ceiling and door sensors mounted on external doors and

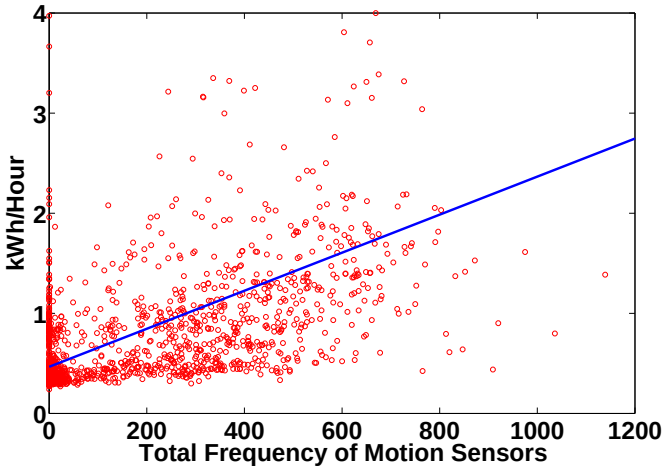
**Table 4.3:** Data features for regression models.

| **Feature Name** | **Description** |
|---|---|
| *Length of day* | the time length since midnight when one instance happens (in seconds). |
| *Day of week* | the current day of week (Monday, Tuesday, Wednesday, Thursday, Friday, Saturday, and Sunday) |
| *Weekday/Weekend* | a binary variable to determine whether the current day is a weekday or weekend |
| *Time of day* | different time slots (morning, noon, afternoon, evening, night, and late night) |
| *Times of individual sensors triggered* | times of different activated motion sensors that were during the time window |
| *Number of kinds of motion* sensors involved | the number of motion sensors that were triggered in the time window. |
| *Total number of times of motion sensor events triggered* | the total number of motion sensor events that were triggered during the time window |

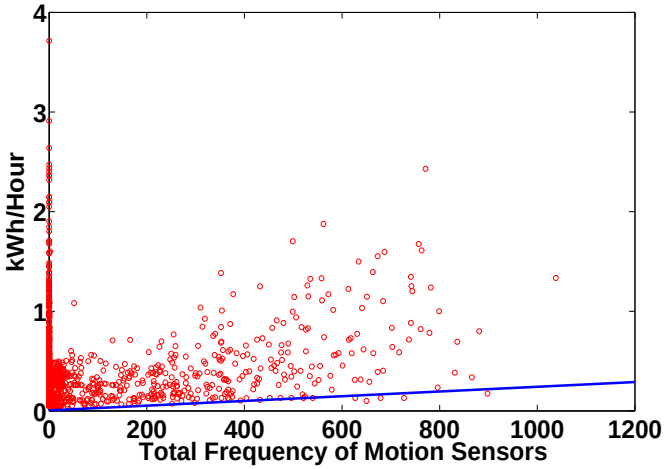
kitchen cabinets. These activities are either directly or indirectly associated with a number of electrical appliances and thus have a unique pattern of power consumption. To illustrate this relationship between behavioral patterns and energy consumption, Figure 4.14 plots the total frequency of motion sensors against energy consumption per hour. Although there is, not unexpectedly, plenty of noise, there is also an obviously linear relationship between motion sensor frequency and power consumption, indicating that the levels of energy consumption grow in direct proportion to users activities at home. We postulate that energy consumption can be measured more accurately given greater sensor density and diversity. In the next section we will further explore different features that can be used to derive predictive models of energy consumption.

### *4.2.2 Feature Selection*

Since a large number of features are generated during feature extraction, it is necessary to determine which features are the most important factors to determine energy prediction. To identify these features, we employ a heuristic minimum-Redundancy-Maximum-Relevance (mRMR) [Peng et al., 2005] selection framework, which selects features mutually far away from each other that still maintain high relevance to the final target. In mRMR, Max-Relevance is used to determine the mean




(a) Kyoto




(b) Tulum

**Figure 4.14:** Plots of total frequency of motion sensors versus energy consumption per hour (Kyoto (a), correlation coefficient: 0.31; Tulum (b), correlation coefficient: 0.24). The blue line shows the least-squares linear fit.

value of all mutual information values between an individual feature $x_i$ and class $c$:

$$I(x_i, y) = \iint p(x_i, y) log \frac{p(x_i, y)}{p(x_i)p(y)} \, \mathrm{d}x \, \mathrm{d}y \tag{4.1}$$

$$maxD(S, c), D = \frac{1}{|S|} \sum_{x_i \in S} I(x_i, c) \tag{4.2}$$

Since the Max-Relevance features may have a very high possibility of redundancy, a minimal redundancy condition can be added to select mutually exclusive features:

$$minR(S), R = \frac{1}{|S|^2} \sum_{x_i, xj \in S} I(x_i, x_j) \tag{4.3}$$

Thus, the operator $\phi(D, R)$ is the mRMR criterion combining the above two constraints:

$$max\phi(D, R), \phi = D - R \tag{4.4}$$

By applying mRMR, the most relevant and least-redundant features among the candidate features are selected. Table 4.4 lists the top six important features from Kyoto and Tulum, respectively, which would be expected to have a large impact on energy consumption. Inspecting the results of feature selection, some specific motion sensor features and time-related features have been selected.

In Kyoto, two sensors from Kitchen/Bathroom are selected and two Bathroom

**Table 4.4:** Selected features using mRMR

| **Max-Relevance, and Min-Redundancy Features** | |
|---|---|
| **Kyoto** | **Tulum** |
| M17 (Kitchen sensor) | M15 (Bath room sensor) |
| M38 (Bath room sensor) | M30 (Bedroom2 sensor) |
| M8 (Living room sensor) | Day of week |
| Length of day | M13 (Bath room sensor) |
| M32 (Bedroom2 sensor) | M22 (Bedroom1 sensor) |
| M18 (Kitchen sensor) | Time of day |

sensors also are chosen in Tulum. It makes sense that residents staying in the Kitchen are very likely to do some cooking, while they may consume plenty of hot water to have a shower in the bathroom. All of these activities are very likely to generate specific energy patterns. The sensors in the Living/Bedroom are also selected. The participants movement patterns throughout the space are either directly or indirectly associated with a number of electrical appliances and thus have a unique pattern of power consumption. For example, when the residents perform the activity in the bedroom, they may turn on a light and operate the computer. It should be mentioned that some time-based features (Length of day, Day of week, and Time of day) are

selected, indicating that the residents may generate various energy usage patterns during different time periods of a single day or days during a week. Thus, those features can also be selected as important factors for energy prediction.

### *4.2.3 Predictive Models*

In this section, we present machine learning techniques for predicting energy consumption based upon sensor features. It should be noted that we are not planning to predict energy consumption very precisely, since energy usage mainly depends on residential preferences on using electrical appliances in the houses, which is hard to capture with our current sensors. The purpose of this section focuses on providing users with basic information about the distribution of their energy consumption based upon their personalized behaviors. In this section, we focus on using real-valued regression for predicting just the total energy consumption. Generally, regression-based learning models have the following form:

$$y = f(x) + \varepsilon \tag{4.5}$$

Here $y$ denotes the predicted energy consumption, x denotes a vector of known features described above, and $\varepsilon$ denotes a zero-mean error term. The goal of our task is to find a function $f(x)$ that has at most  deviation from the actually energy usage

for all the training data. Two well-known machine learning models were leveraged into this work: a linear regression model and a support vector machine. Additionally, since there may be some concerns of known features that may potentially follow non-linear function, a non-linear kernel function is considered to be applied into support vector machine regression models discussed later.

**Linear Regression Model**

A linear regression model [Weisberg, 2005] is applied to model the linear relationship between one or more input features $x \in R^n$ and a target variable $y$. The linear regression model

$$y = \beta^T X + \varepsilon \tag{4.6}$$

for parameters $\beta \in R^n$, and the error term is modeled by a Gaussian distribution. Given a data set, the maximum likelihood estimates of  can be calculated using the least squares method. For linear regression models, the properties of the relevant estimators are easier to determine comparing to nonlinear models. However, there may exist a nonlinear relationship between known features and energy consumption, a support vector machine with a nonlinear kernel will be considered in the next part for exploring such a nonlinear relationship.

**Support Vector Machine Regression Model**

A support vector machine (SVM) model [Gunn, 1998] is an optimal algorithm, which maximizes the margin between training examples and class boundary. It can be applied into both classification and regression problems. In the regression case, SVM estimates the model parameters by minimizing the risk, which is measured using Vapniks $\varepsilon$-insensitive loss function. Given a training data set, a SVM regression (SVR) function can be described as follows:

$$y = w\phi(X) + b \tag{4.7}$$

Here w and b represent the estimator of SVM model, $\phi(X)$ is a map from the original data space of $X$ to a high-dimensional feature space. To obtain the parameters of and $b$, the SVR function can be solved by the following constrained optimization problem by introducing the positive slack variable $\xi_i$ and $\xi_i^*$ as follows:

$$\begin{aligned} &minimize \quad \tfrac{1}{2}||w||^2 + C\sum_{i=1}^{l}(\xi_i + \xi_i^*) \\ &subject\ to \quad \begin{cases} y_i - w\phi(X) - b \le \varepsilon + \xi_i \\ w\phi(X) + b - y_i \le \varepsilon + \xi_i^* \\ \xi_i, \xi_i^* \ge 0 \end{cases} \end{aligned} \tag{4.8}$$

The target function can be computed by

$$f(x) = \sum_{i=1}^{l}(a_i - a_i^*)K(x_i, x_j) + b \tag{4.9}$$

Where $K(x_i, x_j) = (x_i) * (y_i)$ is defined as the kernel function. The main purpose of kernel functions is to deal with non-linear feature spaces without calculating the map $\phi(x)$ explicitly. For our task, we use a linear kernel $K(x, y) = x * y$ and a non-linear kernel $K(x, y) = (x * y + 1)^2$, which are applied to explore a linear and non-linear relationship between known features and energy consumption respectively. Sequential Minimal Optimization (SMO) [Platt, 1998] is reported to be an effective method for improving scaling of the training set and computation time of the SVMs. In our study, the SMO is applied to training the SVMs for solving energy prediction problem.

### *4.2.4 Experiment Results*

Two series of experiments were performed for energy prediction in smart environments. The first experiment uses the sensor data collected during two months in the Kyoto testbed. In the second experiment, we collected data for three months from the Tulum testbed. We evaluated the performance of the algorithms based on 10-fold cross validation: the data instances were randomly divided into 10 approximately equal-sized groups. Specifically, we trained the algorithms over nine of the total groups and tested on the remaining one; we repeated the procedure 10 times for each of the 10 groups, and the average error and correlation coefficient over all

**Table 4.5:** Cross validation performance of different algorithms and time windows for the Kyoto data set. (Item in bold indicates the best performing method)

| Time Window | Linear Regression | | SVM (Linear) | | SVM (Non-Linear | |
|---|---|---|---|---|---|---|
| | Correlation Coefficient | RMSE | Correlation Coefficient | RMSE | Correlation Coefficient | RMSE |
| 1-hour | 0.741 | 0.413 | 0.748 | 0.418 | 0.560 | 0.701 |
| 4-hour | 0.783 | 1.008 | 0.793 | 1.018 | 0.722 | 1.264 |
| 6-hour | 0.785 | 1.796 | **0.838** | 1.519 | 0.741 | 2.077 |
| 8-hour | 0.693 | 1.801 | 0.757 | 1.609 | 0.563 | 2.397 |
| 1-day | 0.562 | 4.88 | 0.701 | 3.978 | 0.723 | 3.796 |

**Table 4.6:** Cross validation performance of different algorithms and time windows for the Tulum data set. (Item in bold indicates the best performing method)

| Time Window | Linear Regression | | SVM (Linear) | | SVM (Non-Linear | |
|---|---|---|---|---|---|---|
| | Correlation Coefficient | RMSE | Correlation Coefficient | RMSE | Correlation Coefficient | RMSE |
| 1-hour | 0.377 | 0.315 | 0.342 | 0.326 | 0.246 | 0.365 |
| 4-hour | 0.598 | 0.666 | 0.605 | 0.654 | 0.466 | 0.893 |
| 6-hour | 0.505 | 0.871 | 0.437 | 0.912 | 0.268 | 1.496 |
| 8-hour | 0.614 | 1.020 | **0.647** | 0.971 | 0.407 | 1.590 |
| 1-day | 0.018 | 7.765 | -0.023 | 9.386 | 0.129 | 38.352 |

the instances were reported. Tables 4.5 and 4.6 show the performance of the alternative algorithms on two data sets. The algorithms were evaluated by two metrics: (1) correlation coefficient, which measures how a regression model fits the data sets; (2) root mean squared error (RMSE) on the energy consumption. Five different time windows were selected for testing the performance on various time scales.

In the Kyoto data set, as seen in Table 4.5, the SVM with a linear kernel obtains the best overall performance with respect both to the correlation coefficient and RMSE. The simple linear regression model is only marginally worse than the linear SVM, which performs much better than the SVM with the non-linear method. For the Tulum training data, the performance between the linear regression model and the linear SVM is very close. Both of these perform better than the non-linear SVM method. We argue that the linear models are preferable for energy prediction in smart environments. These results also validate our hypotheses that the intensity of residents behaviours has a strong linear relationship with energy consumption in houses.

When comparing different time windows, we observe that the Kyoto data instances with 6-hour time windows can be best fitted by the predictive models. The regression models fit best under the 8-hour window for the Tulum data set. It should be noticed that all three regression algorithms are not capable of fitting the Tulum data instances using the 1-day scale. One possible reason is that the residents in Tu-

lum have very similar behaviour patterns at a one-day scale, which is hard to capture with the regression models. By comparing the overall performance between Kyoto and Tulum, energy consumption in the Kyoto environment can be predicted better by the models. We checked the raw power values of the Kyoto and Tulum data sets, respectively. We find that the wave shape of power values in Kyoto fluctuates dramatically. On the contrary, the Tulum energy usage keeps very smooth and steady under different intensity of residential behaviours. That is because there are more and larger electrical appliances installed in Kyoto, such as an air conditioner and a water heater, which may consume much more energy.

Due to the different time window sizes, it is hard to compare the RMSE of the models. Thus, we normalized the RMSE of all time windows to a 1-hour scale RMSE as shown in Figure 4.15. From the figure, for both Kyoto and Tulum, we see that the normalized RMSEs continue to decline dramatically as the size of time window increases except the Tulum one-day window as discussed above. It is also interesting that the normalized RMSE in Tulum is overall lower than the Kyoto normalized RMSE although the Kyoto data can be fit much better by the models based on the correlation coefficient. One possible explanation is that the Kyoto residents consume more energy at unusual times.

The last component of our system is a behavior-based feedback tool to promote energy sustainability in everyday environments. We focus on a pervasive approach

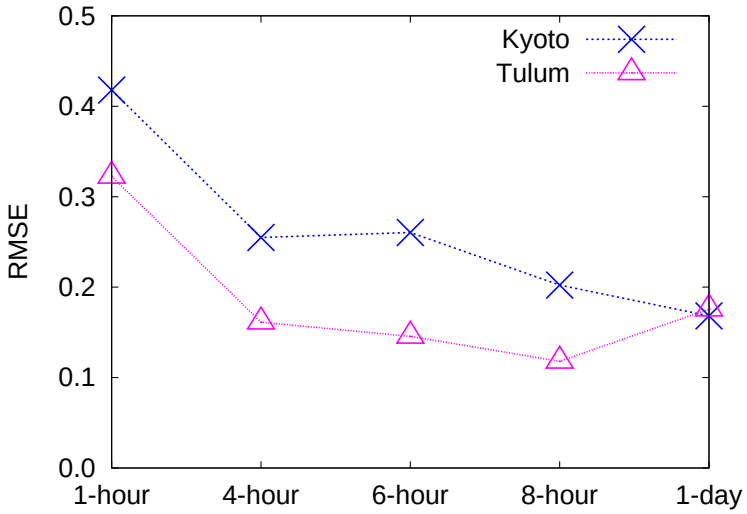
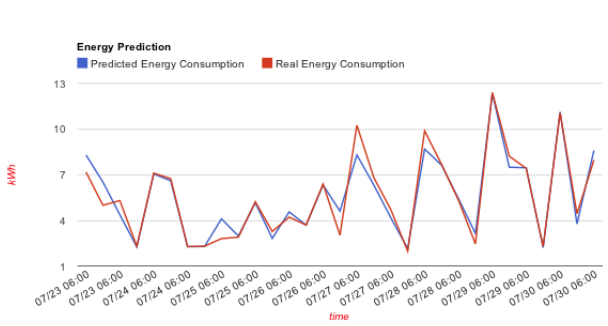
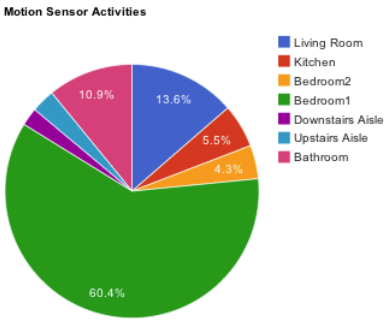



**Figure 4.15:** Comparison of normalized RMSE on different time windows under the SVM with linear kernel.

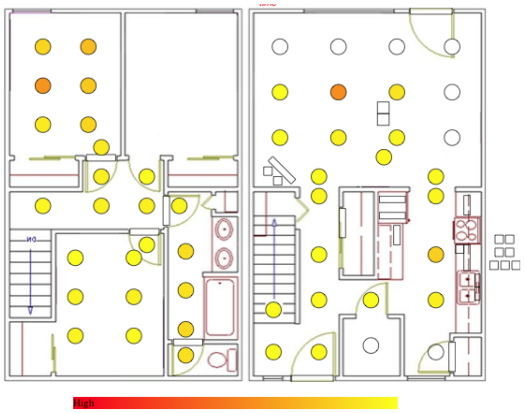
to promote sustainability behavior. Based on regression models described in the previous sections, an end-user application is developed as a web-based solution and can therefore run on a computer display or a mobile device. Figure 4.16(a) shows a line chart for letting residents view and compare predicted energy usage generated by our prediction models and true energy usage in real time. An accompanying activity map, a bar chart, and a frequency chart as shown in Figures 4.16(b), 4.16(c), and 4.16(d), also allow users to quickly determine their energy consumption and the corresponding activities in the environment that impact their energy utilization. A web demo of this tool is available online at `http://demo.ailab.wsu.edu/pv/power.html`. Since our approaches based upon user preferences are fully data-driven, our models are applied particularly to the residents living in our smart environments. Without loss of generality, our models can also be applied to other similar environments.



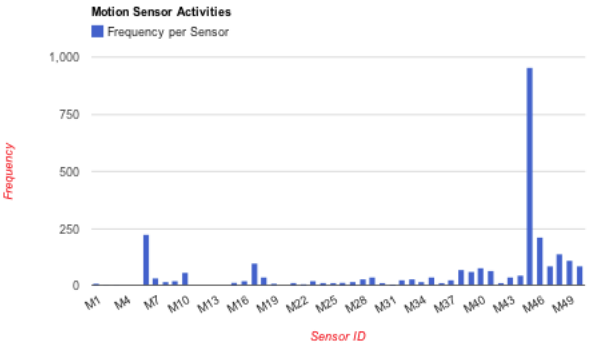

(a) Line chart comparing predicted and real energy usage




(b) Bar chart of sensor activities in each room




(c) Activity map of motion sensors




(d) Frequency chart of motion sensors

**Figure 4.16:** Screenshots of our CASASviz sysem.

# CHAPTER 5. ENERGY ANOMALY DETECTION

In the previous chapter, we discussed home energy consumption models can be built relying on behavioral knowledge. However, these models can do nothing to identify patterns and anomalies based on energy consumption data alone, which can be a good way to explain higher energy consumption. In this chapter, three traditional statistical methods are employed to analyze trends and look for anomalies in energy data. Traditional statistically-based detection algorithms can only identify extreme individual power data points. In contrast, structural power patterns may be more important for the end-users. A pattern-based algorithm anomaly detection will be introduced to identify unusual patterns in the raw energy data.

## 5.1 Statistical Methods for Detecting Anomalies

In this section, the energy data generated by smart environment residents is modeled as a random process with corresponding mean and variation. Here, we make use of three different statistical methods to automatically detect and analyze energy data anomalies and trends in smart home environments. These statistical approaches are: box plot, chart, and CUSUM chart. We test these three methods on energy data

collected in the Kyoto testbed.

**Box Plot**

The box plot [Tukey, 1977] is a quick graphic approach for examing one or more sets of data. A box plot usually displays five important parameters describing a set of numeric data: 1) lower fence, 2) lower quartile, 3) median, 4) upper quartile, and 5) upper fence. As shown in Figure 5.1, the box plot is constructed by drawing a rectangle between the upper and lower quartiles with a solid line drawn across the box to locate the median. The lower and upper fences exist at the boundary of the solid line. The advantage of the boxplot is that it can display the differences between populations without making any assumptions about the underlying statistical distribution. In addition, the distance between the different parts of the box helps indicate the degree of spread and skewness in the data set.

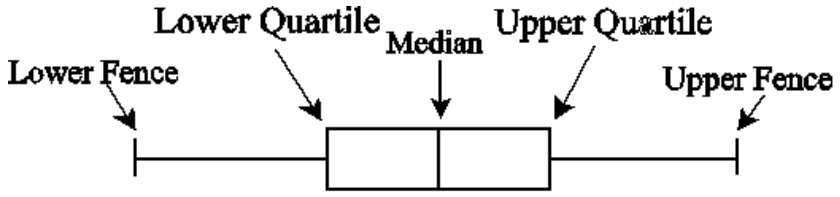



**Figure 5.1:** Configuration of a box plot.

Statistical Process Control (SPC) [Deming, 1975] is the application of statistical charting techniques for detecting shifts in mean or variability of a process. Here, energy usage data will be modeled as a random process whose mean and variance

could be estimated by the sample data. We will utilize two SPC techniques to identify abnormal energy usage data as follows.

### $\overline{X}$ Control Chart

The first technique focuses on generating control charts. In statistical process control, control charts are particularly useful for monitoring quality and giving early warnings that a process may be going out of control. A typical control chart has control limits set at values such that if the process is in control, nearly all points will lie between the upper control limit (UCL) and lower control limit (LCL). Assume that for an in-control process, the data collection $X$ follows a normal distribution with mean value, $\mu$, and stand deviation, $\sigma$. If $\overline{X}$ denotes the sample mean for a random sample of size $n$ selected at a particular time, the $\overline{x}$ chart for determining control limits first calculates the mean $E(\overline{X}) = \mu$ and standard deviation $\sigma_x = \sigma/\sqrt{n}$ of the sample values. Next, upper and lower controllimits are defined as $(\mu+3\sigma_x/\sqrt{n}, \mu-3\sigma_x/\sqrt{n})$. These control limits can be used to identify the outliers in energy data that occur in the specific monitoring time window. The plot of mean values associated with the control limits are used to determine when the process is "out of control". In the case of energy data analysis, when an important acute change has occurred, the $\overline{x}$ chart can identified the location of this change. The disadvantage of a $\overline{x}$ control chart is its inability to detect a relatively small change in a process mean because the ability

to judge the process as being out of control at a particular time depends only on the sample at that time, and not the past history of the process.

**CUSUM Control Charts**

Cumulative sum (CUSUM) control charts [Kafadar, 2003] have been designed to address this problem. The CUSUM chart works as follows: Let $\mu_0$ denote a target value or goal for the process mean. The cumulative sums can then be calculated using Equation 5.1.

$$S_n = \sum_{i=1}^{n} (\overline{x} - \mu_0) \tag{5.1}$$

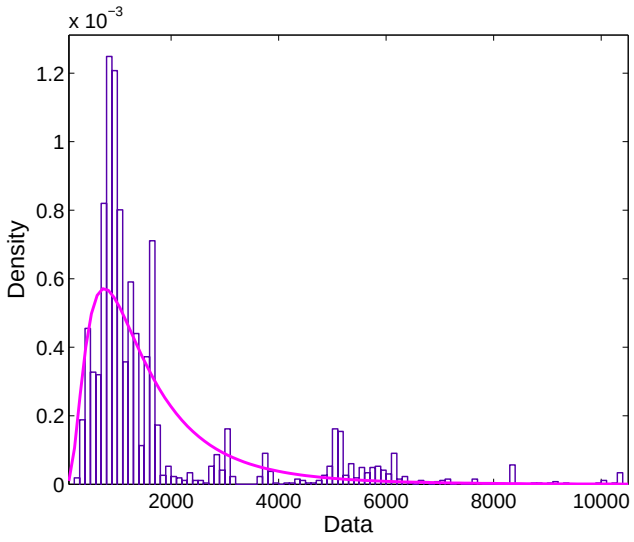
These cumulative sums are plotted over various time windows and a V-shaped mask is superimposed on the graph of the cumulative sums. At any given time, the process is judged to be out of control (or anomalous) if any of the plotted points lies outside the V-mask, either above the upper arm or below the lower arm. An out-of-control situation has been identified by the V-mask because one point in the time window lies above the upper arm. The V-mask is calculated based on the lead distance d and the rise distance h. The parameter-defined variations in the shape of the V-mask will thus affect the type and number of outliers that are detected.

### *5.1.1 Experiment Results*

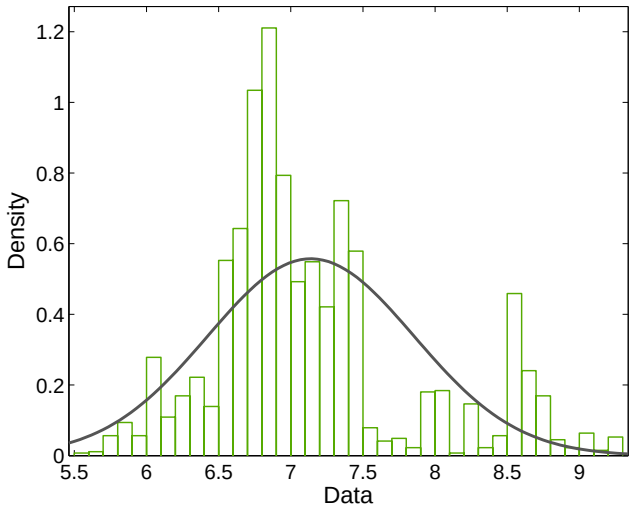
We performed two experiments using the energy data collected during an entire year in our CASAS smart apartment testbed. The first experiment detects abnormal energy wattage during any single day. The second experiment looks for novelties in energy Kwh data consumed each week over the course of the entire year.

When we generate a $\overline{x}$ control chart, there is an assumption that the random process follows a normal distribution. Thus, we need to examine whether the energy data during different time windows fits the normal distribution. Based on the Central Limit Theorem [Rice, 2006], if a random sample of $n$ observations is selected from any population, the sampling distribution will be approximately normal. Unfortunately, the energy data for different time granularities in our smart home environment often demonstrate a positive skew. Thus, we use the lognormal distribution to describe the energy data distribution $x$. In this case, $ln(x)$ should follow a normal distribution approximately. As shown in Figures 5.2, 5.3 and 5.4, the plots on the left show how the original energy data $x$ fits the lognormal distribution and the plots on the right describe how the normal curve simulates the variation in log energy values. From the graphs, we see that the log of the energy data can basically fit the normal distribution very well. Thus, we can continue to use $\overline{x}$ charts for detecting energy data outliers.

For the first experiment we focus on energy wattage data collected for one day

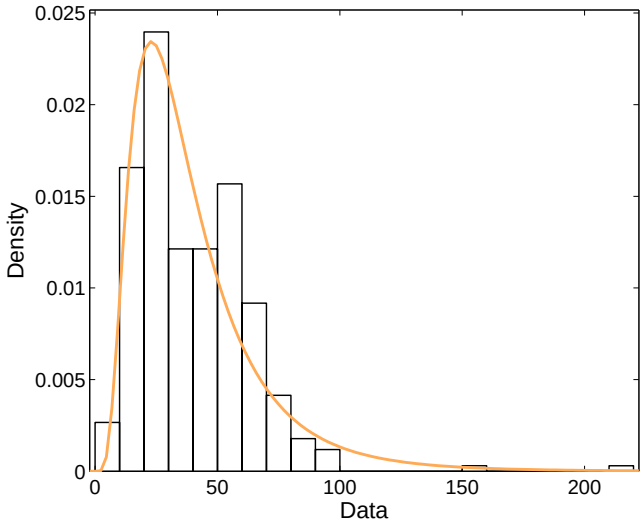



(a) Raw Energy Data

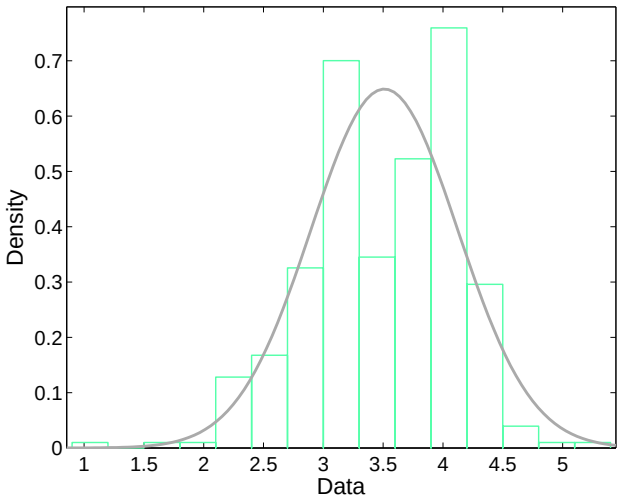



(b) Log-scale Energy Data

**Figure 5.2:** The lognormal and normal distribution of energy data (W) for one day.

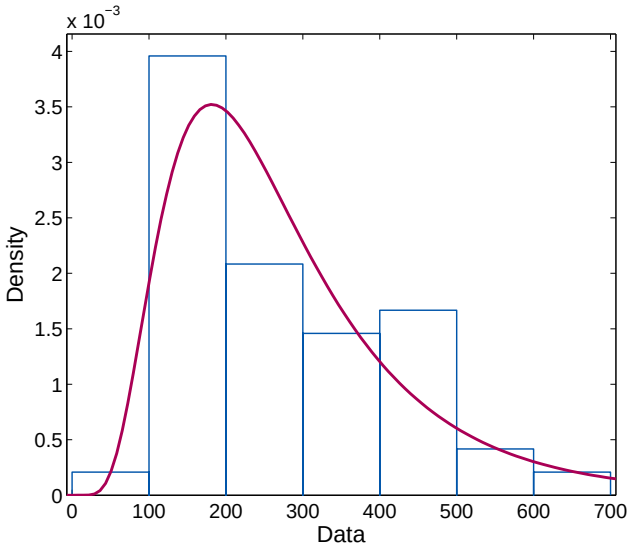



(a) Raw Energy Data

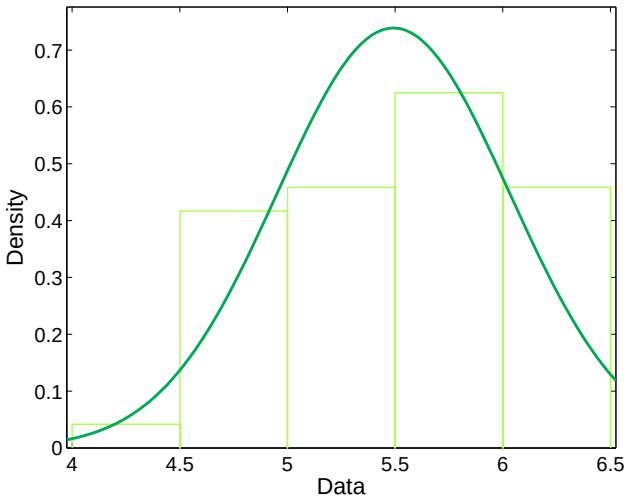



(b) Log-scale Energy Data

**Figure 5.3:** The lognormal and normal distribution of energy data (kWh) for one day.

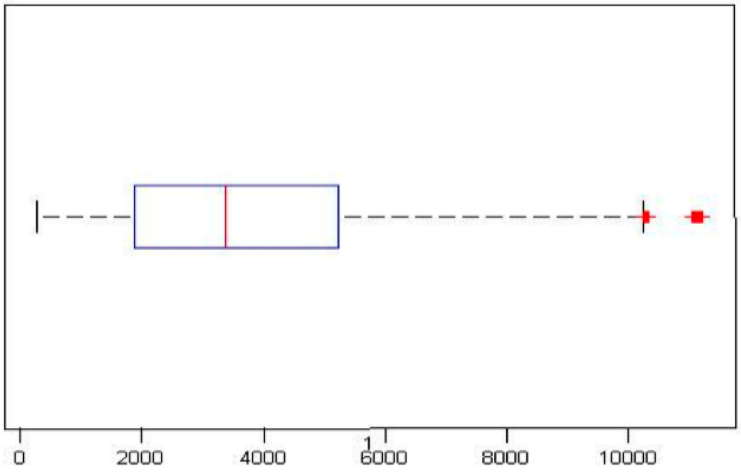



(a) Raw Energy Data

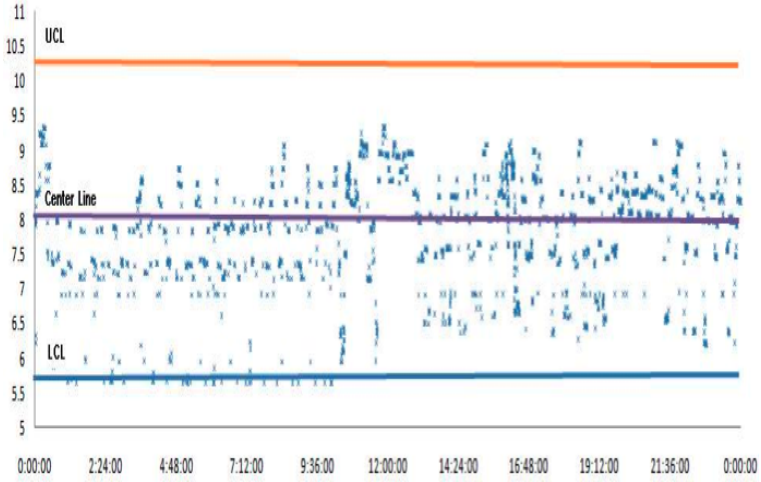



(b) Log-scale Energy Data

**Figure 5.4:** The lognormal and normal distribution of energy data (kWh) for one week.




(a) Box plot chart




(b) $\overline{x}$ chart

**Figure 5.5:** Bot plot chart (left) and $\overline{x}$ chart (right) of energy data (in Kwh) for one day.

in the smart environment testbed. The purpose of this experiment is to detect the energy data outliers and determine possible reasons for the anomalies. Figure 5.5(a) shows the box plot graph of the data. The points located on the right side represent the outliers. We examined those outliers in detail and found out these abnormal readings occur during two main time intervals. The first set of anomalies was mainly concentrated at around midnight. One reasonable explanation is that all the heaters in our smart home worked at the same time because the temperature of that time is the lowest during the day. The outliers in the second group are located at the middle time of the day, which is residents cooking time and the large appliances are being used for cooking such as the microwave, the stove and the oven, all of which would give rise to dramatically increasing energy consumption. For the $\overline{x}$ control chart shown in Figure 5.5(b), all the outliers fall below the lower control limit. All of these anomalies occurred between 01:00 am and 06:00 am, which is the common sleep time for the residents. most of the appliances are idle during that time interval.

The CUSUM chart as described in Figure 5.6 detects some outliers not detected in the previous experiment because the CUSUM chart is very effective for small shifts and the process can be judged out of control depending on the past history of the process. However, a drawback of the CUSUM chart is that it is relatively slow to respond to large shifts and some special data patterns are also hard to analyze and explain. In this experiment, the CUSUM chart highlights a large number of outliers,

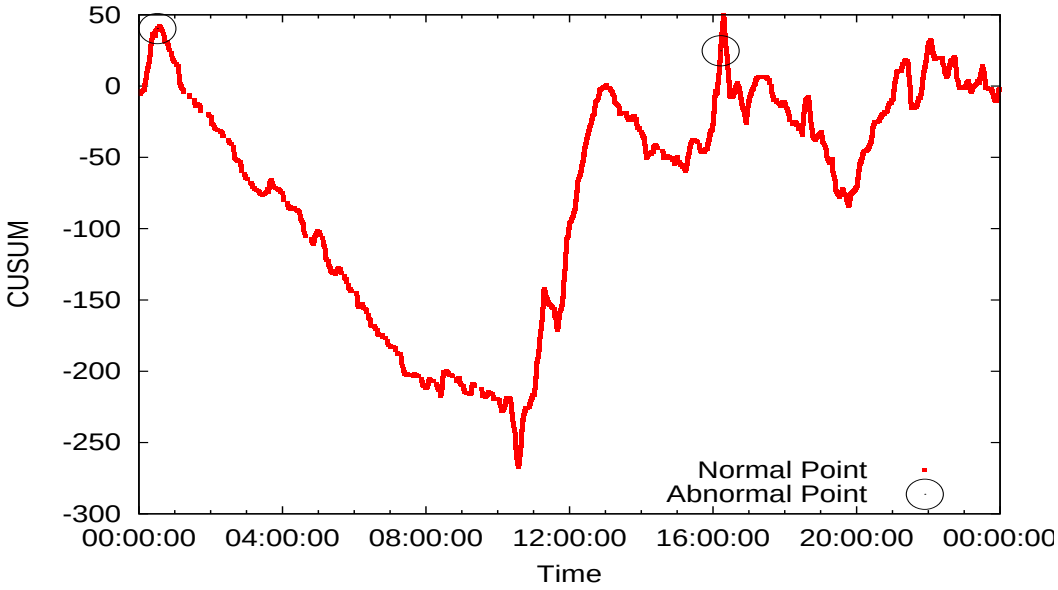
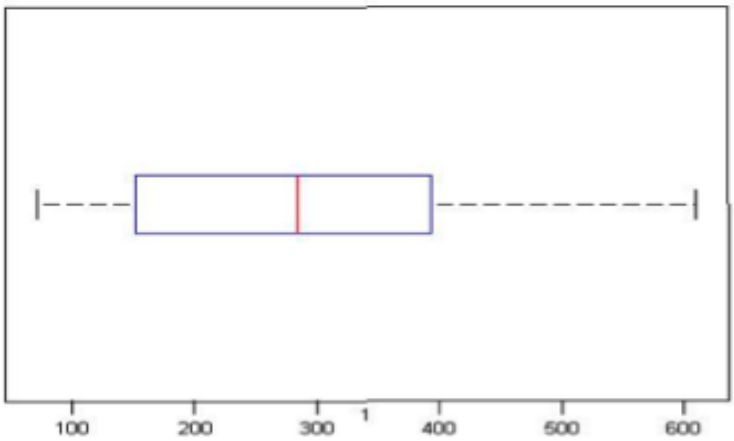


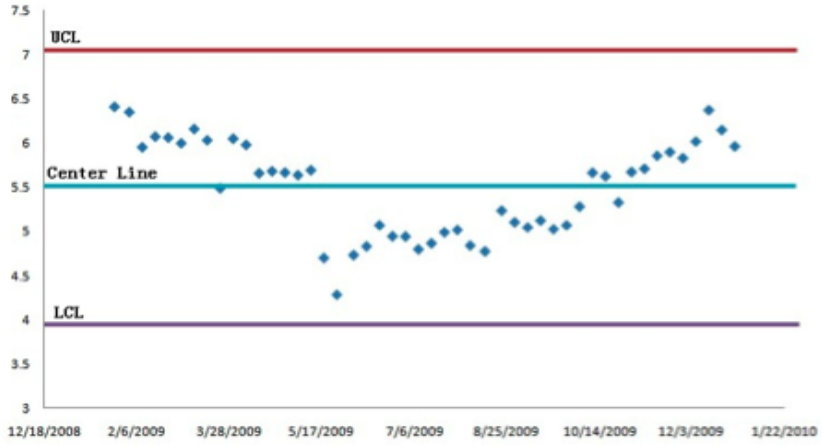

**Figure 5.6:** A CUSUM chart of energy wattage data for one day (d = 145.1, h = 30).

many of which are difficult to explain. However, these results provide some valuable information for understanding human behaviour in the smart apartment. Novelties were detected at times 00:31:07 and 16:12:50, both of which represent turning points in energy usage during the day. After these times, energy consumption decreased continuously, perhaps because some large electrical devices were turned off.

The second experiment analyzes energy consumption data (Kwh) by week over a year timeframe in order to look for the long term trends of energy usage and its relationship with other relevant elements like weather variation. However, from Figure 5.7, none of the outliers can be detected by the box plot or $\overline{x}$ control charts. Thus,




(a) Box plot chart




(b) $\overline{x}$ chart

**Figure 5.7:** Bot plot chart (left) and $\overline{x}$ chart (right) of energy data (in Kwh) by week for one year.

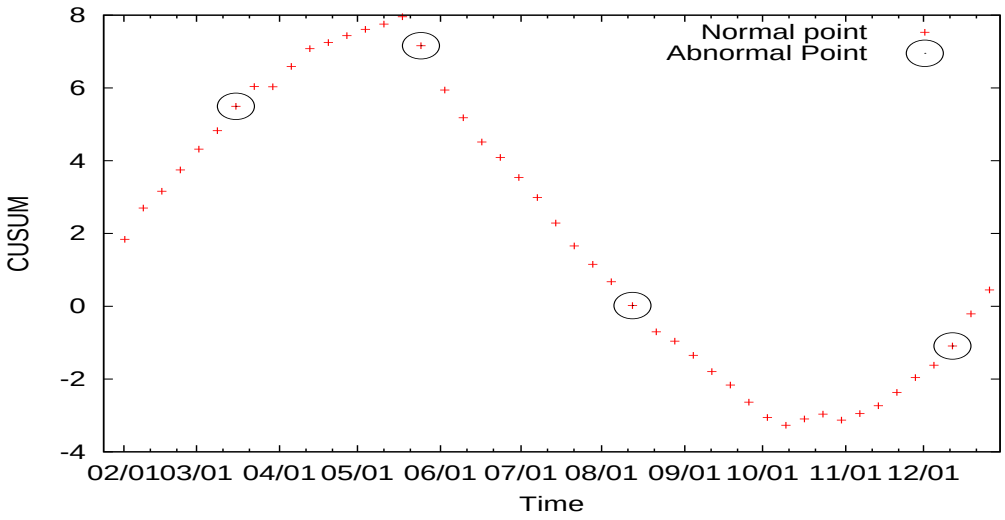


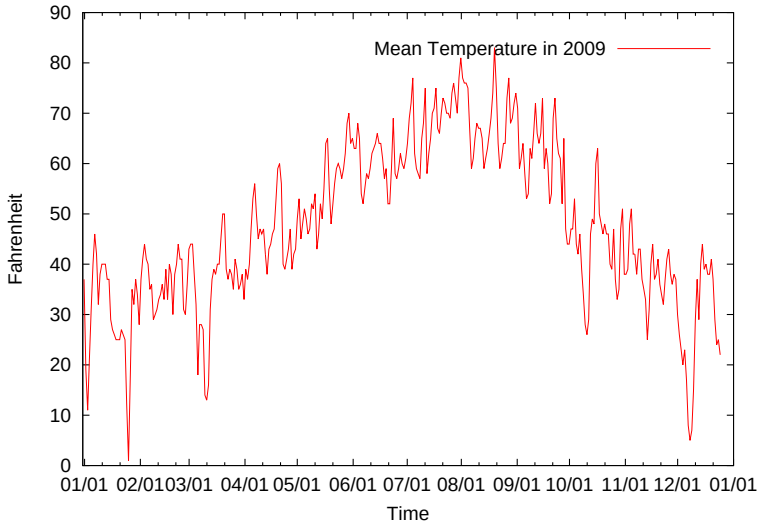

**Figure 5.8:** CUSUM chart of energy data (Kwh) by week during one year (d = 4.59, h = 0.95).

the variation of this dataset does not experience extreme changes during the year.

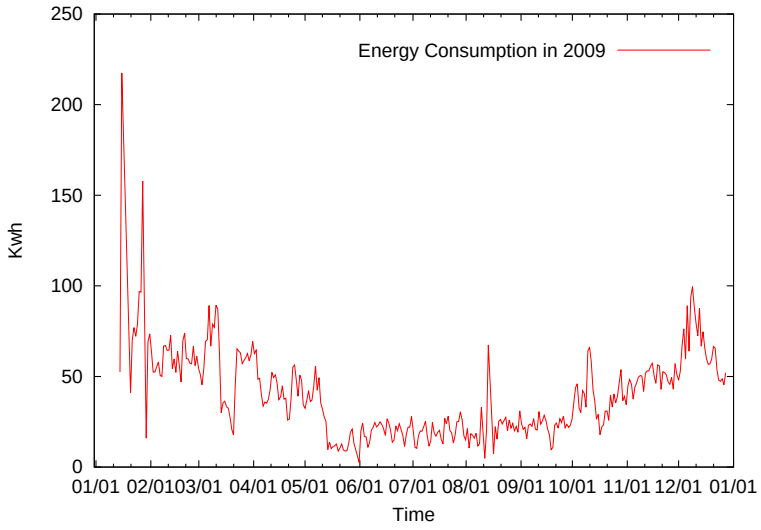
As shown in Figure 5.8, the CUSUM chart shows the periodic pattern of the cumulative sum. The CUSUM chart identifies four energy data abnormalities (on the dates 03/16, 05/25, 08/13, and 12/12), which represent changes for the four different seasons (Spring, Summer, August and Winter). That result gives us a clue that there may be a possible strong relationship between seasonal temperature changes and energy usage. Thus, we continue to explore this relationship. Figure 5.9 plots the trends in energy usage and average external temperature. Historic average regional temperature values are obtained online [Weather Underground, 2013]. This figure shows

that there exists a strong relationship between energy usage and external temperatures during the same time. When the temperature increases or decreases, the energy usage consumed by the residents will decrease or increase correspondingly most of the time. The likeliest reason is that the heaters in our smart home environment will consume different amounts of energy with temperature changes. In the winter, the heater can consume more energy with the temperature decreases. Due to the specific location of our smart environment, the temperature in the summer is not very high. Thus, the residents do not need to use air conditioner in the apartment. That is why the energy consumption does not exhibit a large variation in the summer. In our testbed, the heaters are key influences on energy efficiency. In the future, residents might be able to utilize other heating sources by opening blinds or decreasing the temperature at night in order to improve energy efficiency in the apartment.

Analyzing the results of our experiments, we see that statistical approaches can be useful for detecting and identifying anomalies in energy usage, which in turn provides insights on human behaviour and gives the residents some valuable insights with which they can improve their own daily patterns to reduce energy usage. However, there are also some drawbacks of these methods. Box plots and charts only detect relatively large changes in a process mean and sometimes fail to detect small changes. This is why both of these methods did not detect the outliers for the weekly energy data. In contrast, CUSUM charts can be very effective for small shifts based on the




(a) Mean Temperature Chart




(b) Energy Usage Chart

**Figure 5.9:** The comparison between mean temperature (top) and energy usage (bottom).

past history of the process. However, CUSUM charts are relatively slow to respond to large changes.

## 5.2 Pattern Clustering Detection

Statistical methods are very useful to detect extreme energy values. However, several continuous structural power values can represent power patterns much better. To address this issue, we analyze normal patterns by clustering sequences of power usage values. This analysis is useful because the cluster descriptions can provide users with insights on their daily habits and resource usage as well as provide software algorithms with a model of normal usage in a particular environment. At the same time, the clusters provide a baseline against which anomalies in energy usage can be identified. Anomaly detection is valuable because the anomaly may indicate an unnecessary use of resources (e.g., an appliance was accidentally left on), an unsafe state, or possibly noise in the dataset, which needs to be removed.

The amount of real time instantaneous power consumption is recorded in our smart homes. In our method, this data is first discretized into $k$ value ranges using equal-width binning [Liu et al., 2002] and then is convert to symbols. Through binning, an energy sensor sequence E can be transformed into a new energy symbol sequence $S$, which is defined as:

**Definition 1.** *An **energy symbol sequence** $S = s_1 s_2 \dots s_n$ is an ordered set of $n$ symbol variables over the alphabet $\Sigma$, where $\Sigma = \{a, b, c, \dots\}$ and $\|\Sigma\|$ is equal to the number of bins $k$. All energy values in the range for the $i^{th}$ bin are represented by symbol $i$ in the sequence.*

After converting raw power data into a symbol sequence, the algorithm discovers patterns in energy usage data by employing suffix trees [Gusfield, 1997]. Unlike other data mining methods, which are exponential in their complexity, this algorithm can generate a suffix tree in $O(n)$ time for a symbol sequence of length $n$, and spend $O(m)$ time to search for a subsequence of length $m$, regardless of $n$. A formal definition of this tree follows.

**Definition 2.** *Given a string $S'$ over the alphabet $\Sigma$ and a unique termination character $\$ \notin \Sigma$, the string resulting from appending $\$$ to $S'$ can be defined as $S = S'\$$. Let $|S| = n$ and $suff(S, i) = S_i S_{i+1} \dots S_{|S|}$ be the suffix of the string $S$ starting at $i^{th}$ position. The **suffix tree** of $S$ is a compacted trie-like data structure that stores all suffixes of a string $S$ over the alphabet $\Sigma$.*

Traditional suffix tree construction algorithms start from the root and follow a unique path matching characters in $suff(S, i)$ one by one until no more matches are possible. If the traversal does not end at an internal node, it creates a new internal node at that location. For a tree with $n$ nodes, the total running time of the

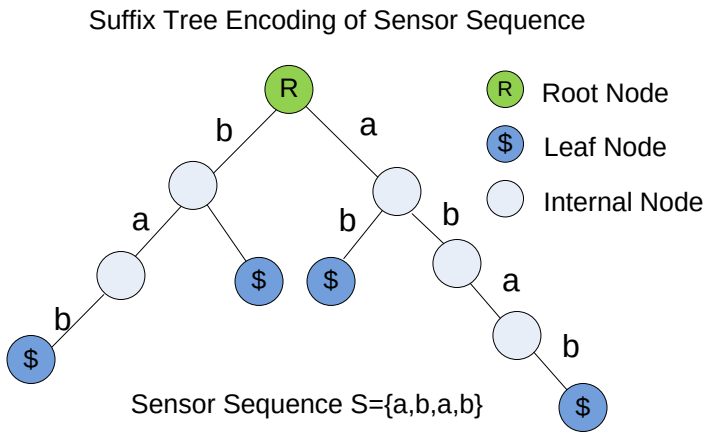



**Figure 5.10:** A suffix tree defined on a symbol sequence $S$ with length $m$ can represent every subsequence in $S$ with at most $2m$ nodes

algorithm is $\sum_{i=1}^{n}(n-i+1) = O(n^2)$. In order to achieve $O(n)$ running time, we use McCreight's algorithm [McCreight, 1976] to construct a suffix tree by applying suffix links to speed up the insertion of a new suffix.

A graphical illustration of the transformation of an energy sequence into its equivalent suffix tree is shown in Figure 5.10. By definition, no two edges emanating from a node in a suffix tree begin with the same symbol, which implies that every unique subsequence in $S$ starting from the root node can be generated by traversing through the suffix tree. We consider these subsequences as **energy patterns**, which are defined as:

**Definition 3.** *Let an* ***energy pattern*** $p_i$ *in* $S$ *represent the subsequence generated by*

**Table 5.1:** Examples of energy patterns

| **Energy Pattern** | **Pattern Length** | **Raw Energy** | **Pattern Frequency** |
|---|---|---|---|
| CC | 2 | 752, 742 | 26952 |
| ZFZ | 3 | 5000, 1021.2, 5007 | 13 |

*traversing a path in the suffix tree, where p represents the sequence of symbols visited along the path and the length of this energy pattern is i. The frequency of an energy pattern $p_i$ in S is denoted by $f(p_i)$, which is equal to the number of the leaf nodes found in the subtree rooted at the end of the subsequence $p_i$.*

Table 5.1 shows two examples of energy patterns and their corresponding frequencies. In the first case, energy readings of 752 and 742 fall in the same bin (value range) and are mapped to symbol C. The sequence of energy readings CC occurs 26,592 times in the data file and thus is a much more common pattern than the one found in the second line of the table. In the context of this brief example sequence CC might be considered a pattern of interest, while sequence ZFZ might be considered an outlier or anomaly.

### *5.2.1 Sequence Clustering*

To detect abnormal situations, we next cluster all the energy patterns into groups with similar patterns and identify sequences that do not fit well in any cluster. Intuitively, for an energy symbol sequence $S$, we consider an energy pattern $p_i$ to be an outlier if this energy pattern is far from the centroid of the cluster.

Cluster analysis is a data mining technique that is often used to identify various groupings or taxonomies in datasets. We apply clustering to power sequence values in order to gain a better understanding of the data, to identify groupings of normal energy usage, and to use as a baseline for identifying abnormal energy usage patterns. A clustering algorithm takes features of the data as input and creates a classification scheme which is represented as a set of disjoint clusters, each of which can be described by a middle point, or cluster centroid.

One important step in our clustering process is to decide a distance measure, which is used to group sequences together in a cluster and should reflect the similarity of two sequences. In this paper, we use a two-step process. We first restrict clusters to contain only patterns of the same length. That is because the suffix tree algorithm naturally groups patterns into distinct lengths, and our algorithm further divides these groups into subgroups using the clustering algorithm. From these groups we next employ Euclidean distance measure, which is a geometric distance in the mul-

tidimensional space and is widely used by clustering algorithms. Based on specific property of energy patterns, we select three related features, which will be used to measure the dissimilarity, or distance between energy patterns.

**Pattern Variance between Energy Patterns.** As defined in Definition 5, $p_i = s_1 s_2 \dots s_i$ is an energy pattern, where $s$ is an energy symbol after binning. The distance between two symbols $|s_x - s_y|$ can be be estimated as the alpha-numeric distance between the symbols. To determine pattern variance, we measure the distance between each corresponding symbol in the pattern. Thus the pattern variance between $p^1$and $p^2$ with length $i$ is defined as:

$$d_1(p^1, p^2) = \sum_{j=1}^{i} |s_j^1 - s_j^2| \tag{5.2}$$

**Within-Pattern Variance**. Because changes in power occur when appliances are switched on or off, the difference between two consecutive symbols in an energy pattern may indicate a change in the status of the appliances. Thus, the variance within this energy pattern captures the usage status of the appliances. The within-pattern variance of an energy pattern $p$ can be calculated as $v_i = \sum_{j=2}^{i} |s_j - s_{j-1}|$ . We define the difference in within-pattern variance between two energy patterns $p^1$ and $p^2$ as:

$$d_2(p^1, p^2) = |v^1 - v^2| \tag{5.3}$$

**Frequency of Energy Pattern**. Another important feature we cannot ignore is the frequency of an energy pattern, as defined in Definition 5. The lower the frequency is, the more likely this pattern is an outlier. If the frequency of a pattern is relatively high, it may represent a normal pattern of usage. The frequency difference between energy patterns $p^1$ and $p^2$ is calculated as:

$$d_3(p^1, p^2) = |f(p^1) - f(p^2)| \tag{5.4}$$

To balance the impact of these three metrics, all these three distance values are normalized to the scale [0, 1] and the final distance between two energy patterns $p^1$ and $p^2$ is estimated as:

$$d(p^1, p^2) = \sqrt{d_1(p^1, p^2)^2 + d_2(p^1, p^2)^2 + d_3(p^1, p^2)^2} \tag{5.5}$$

### *5.2.2 Outlier Detection*

In residential settings, anomalies can be detected over different time scales including single anomalous energy readings or anomalous days, weeks, or months of energy usage. In addition, several types of anomalies can be observed. One type of anomaly is characterized by a energy usage value that is outside of the normal range. When such values are noted for single readings, they may reflect abrupt changes in

usage due to large appliances being turned on or off. They may also indicate errors in the energy usage monitoring mechanism or indicate a more drastic situation such as a blackout. We target these types of "out of range" anomalies with our detection method.

In contrast, anomalies may also occur due to changes in resident behavioral routines. For example, the residents may leave the home on vacation or host a large, extended party. These changes can result in shifts away from normal patterns of usage. These types of unexpected changes would not typically be detected by considering just a typical range of values. We do not address this type of anomaly detection in our current work. However, we target this as an avenue for possible future research.

In the second step of our analysis, we use the generated clusters to identify outliers in the energy usage data. The outliers are defined as energy usage sequences that fall as far as possible from the centroid of any cluster. Detecting these outliers consists of two stages. In the first stage, we cluster the energy sequences and calculate the cluster centroids. In the second stage, we calculate the distance of each energy pattern sequence to the cluster centroids. The greater this distance is, the more likely it is that the pattern is an outlier. Patterns for which the distance is greater than a pre-defined threshold are considered to be outliers and indicate anomalous energy usage.

From this discussion it is apparent that the choice of a threshold value greatly

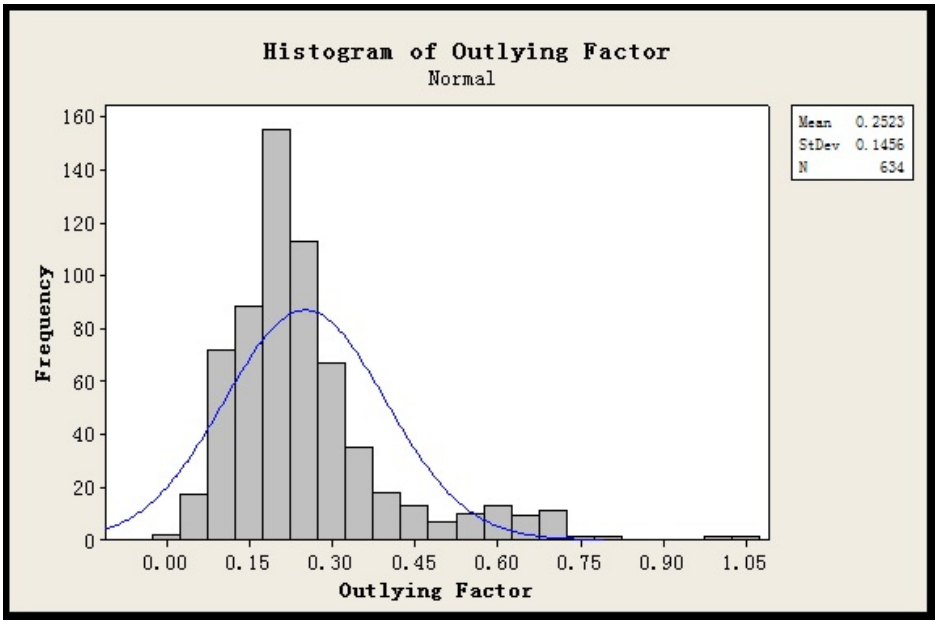



**Figure 5.11:** Histogram of outlying factors of all energy patterns ($k = 50$ bins).

influences the selection of outliers. A variety of methods could be used to select the threshold. These could be based on statistical parameters of the data itself or user-selected parameters such as the rarity of the anomalies that are being reported. To determine the value for this application domain we plot a histogram of all pattern distance values to the centroid (also referred to as outlying factors, see Figure 5.11). It was noted that these outlying factors follow a normal distribution, which means that 99.7% of the patterns will then fall within three standard deviations of the mean. To detect the outliers, we only consider the patterns that fall outside of this area. As an option, the user can manually assign an appropriate value as a threshold for

identifying the outliers.

### *5.2.3 Household Power Data Simulator*

One of the most challenging problems for anomaly detection is that it is difficult to attain ground truth to measure the performance of the algorithm. Obtaining ground truth may require human effort to analyze a vast amount of power data and mark the anomalies manually. This method is time-consuming and error-prone. To deal with this problem, a power data simulator is designed to generate energy data time-series given knowledge of energy fluctuations caused by the status change of appliances. We next provide details of the simulator we have designed for this purpose.

In the first step, the original energy data are first examined. As shown in Figure 5.12, the red line represents the real power data generated by the appliances in the CASAS smart homes. From the figure, it can be seen that the energy consumption will not become zero, even when the major appliances have been turned off. That is because there are some appliances which are always in use at varying levels, such as refrigerator, phone charger, camera etc. The purpose of the first step is to identify this basic level of home energy usage without the major appliances on. Based on real smart home data, this basic home energy level can be simulated using normal distribution

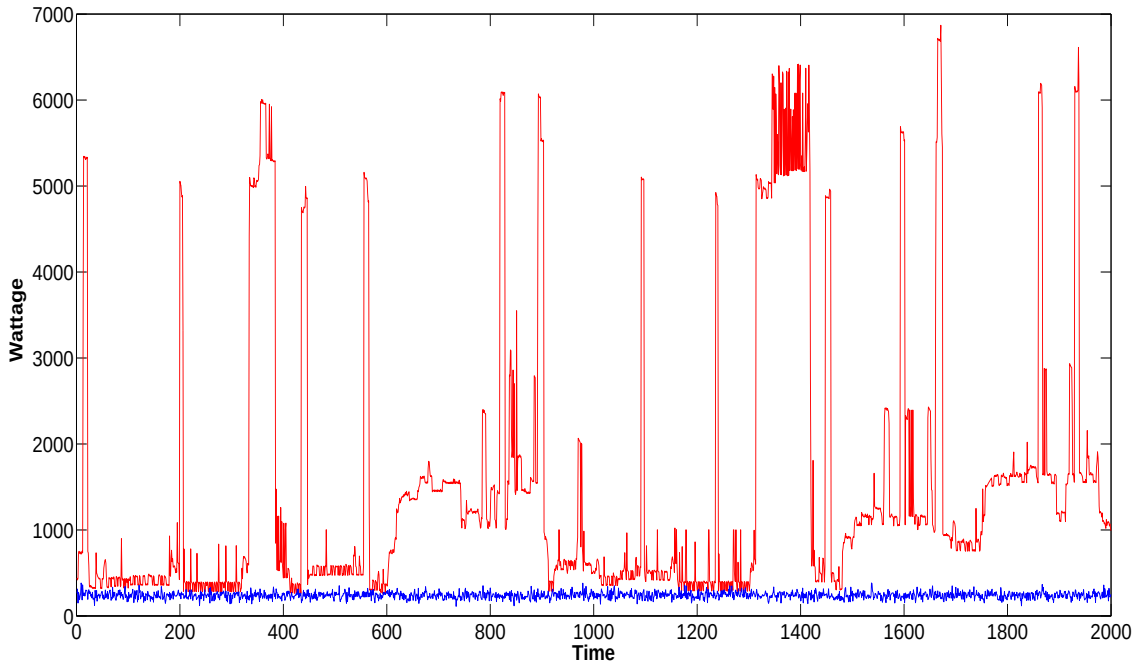


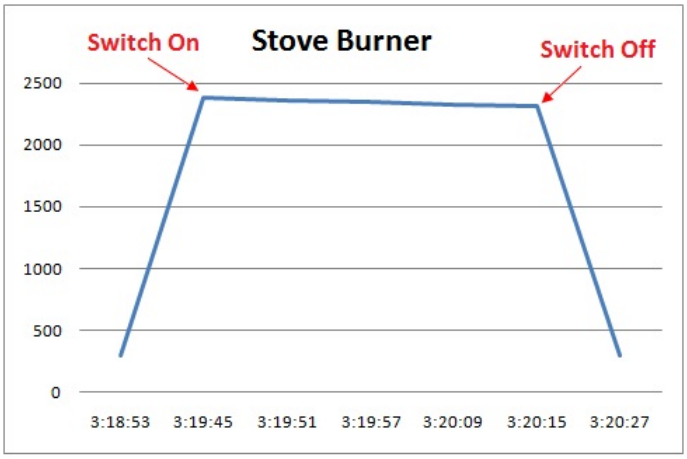

**Figure 5.12:** Extract the power base line from the real power data (red line: real energy data in the CASAS smart home; blue line: the base energy data without any appliances on).

data generator. In Figure 5.12, the bottom of the blue line can be considered as the basic energy data without any major appliance on.

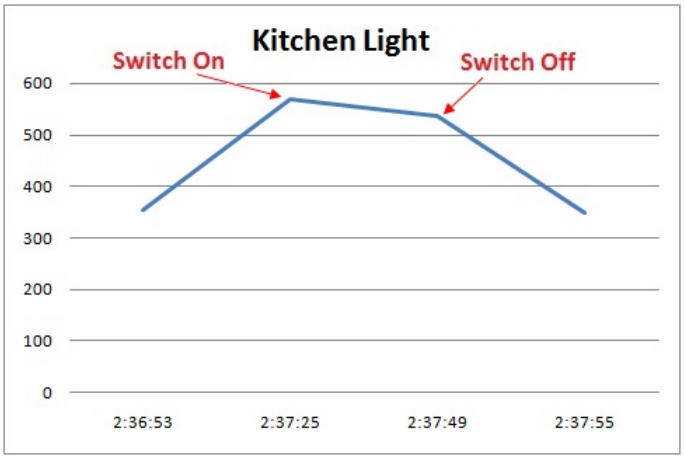
To simulate real energy usage in the homes, the patterns of appliance usage first need to be investigated. Figures 5.13(a) and 5.13(b) show the changes in power that occur when the stove burner and kitchen light are switched on or off. From the figures, we can see that the absolute change in value will vary for each device because each device has a specific rated power. These values can be considered as a valuable key to recognize the switch status of the device though this signature may not be unique for each device. In order to attain accurate energy change values, we repeated switching the device on or off ten times, then recorded the increased or decreased average power consumption. Table 5.2 summarizes the device name and the power mean and variance values when the device is switched on or off.

The second step of our tool is to randomly insert energy change values to the basic energy usage given switch frequency of each appliance. Figure 5.14 dictates the simulated energy data after inserting energy change values of each appliance. The change values of each appliance are generated by normal distribution $N(\mu, \sigma)$, where $\mu$ is the mean of energy change and $\sigma$ is the variance of energy change.

The last step of the simulated tool is to inject the anomalies randomly into the simulated energy data. Moreover, the various levels of the anomalies can be chosen. In our experiment, we took into account this simulated energy data as the




(a) Abrupt power change when switching on and off the stove burner.




(b) Abrupt power change when switching on and off the kitchen light.

**Figure 5.13:** Examples of power change when switching on and off.

**Table 5.2:** Table of device energy changes.

| **Device Name** | **Average Energy Change (Watt)** | **Variance Energy Change (Watt)** |
|---|---|---|
| Living Room Light | 205 | 20 |
| Kitchen Room Light | 205 | 20 |
| Stove Burner(big) | 2010 | 200 |
| Television | 80 | 10 |
| Microwave | 1240 | 100 |

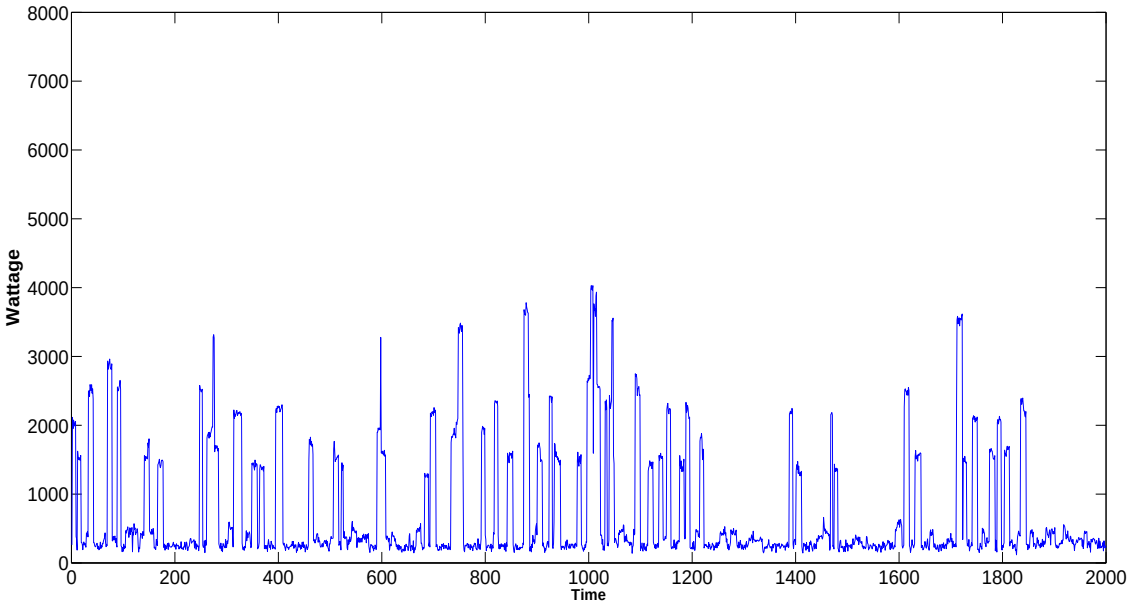



**Figure 5.14:** Simulation of the power fluctuations caused by the appliances.

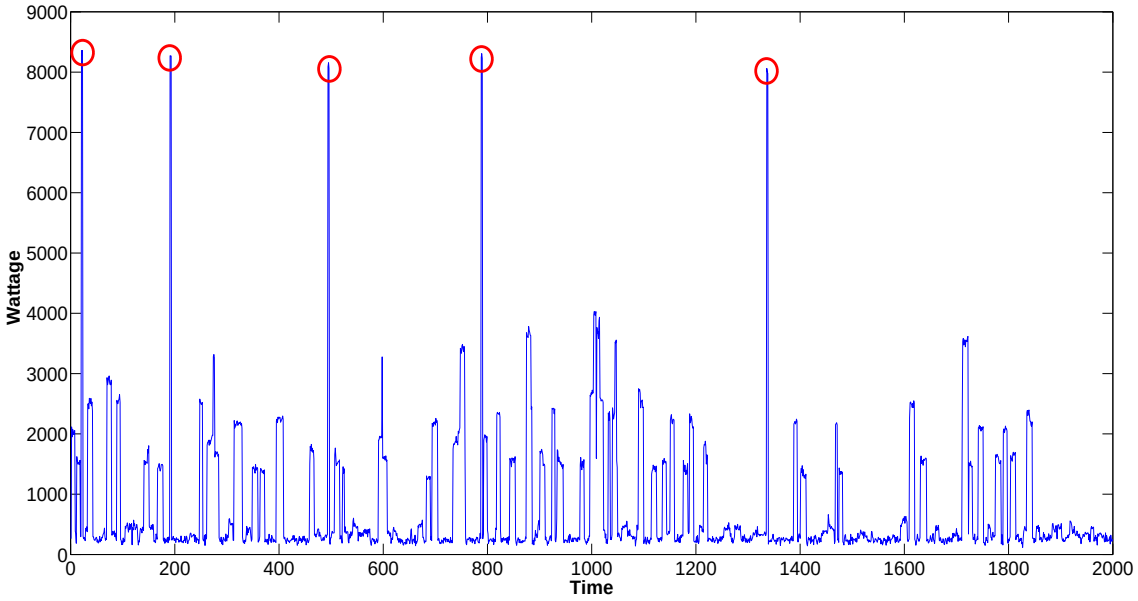



**Figure 5.15:** Insert the anomalies manually into the simulated energy data.

ground truth to measure the performance of our anomalies detection. Figure 5.15 shows the new simulated energy data after injecting the anomalies manually. The red circles mark the anomalies in the simulated energy data. The frequency and severity thresholds of the anomalies can be decided by the users.

### *5.2.4 Experiment Results*

Two experiments were performed for detecting anomalies in energy data. The first group of experiments looks for abnormal energy data during a single week from

two of our smart apartments, Kyoto and Tulum, respectively. In the second group of experiments, four methods are employed to detect outliers in the simulated energy data with the injected anomalies. In comparison, the proposed pattern-based approach was applied to detect the outliers existing in the same dataset.

The first group of experiments focuses on energy usage data collected for a single week in the Kyoto and Tulum testbeds. The threshold is generated by three standard variance beyond the mean. The purpose of this experiment was to detect energy outliers and determine possible reasons for these outliers. To explore potential reasons for the anomalous usage patterns, those outliers were examined in detail and referenced with knowledge of the residents' behavioral routines. It was discovered that these abnormal events represent two types of occurrences. The first set of outliers was mainly due to large changes in energy usage, when the residents had sustained high-level energy consumption over a long time. Some of the big appliances, including the water heater, consume more energy than others and can create anomalies when there are long showers. In addition, during the middle of the day the residents often do their cooking and large appliances are being used for cooking such as the microwave, the stove and the oven, all of which would give rise to a dramatic increase in energy consumption. To respond to these outliers, the residents can analyze their energy needs during these activities to identify energy-saving behaviors.

The outliers in the second set consisted of two successive energy events, whose

**Table 5.3:** Example of an outlier.

| | | |
|---|---|---|
| 2009-06-01 23:31:02 | P001 | 1001.5 |
| 2009-06-01 23:31:02 | P001 | 356 |

values are different but occur at the same time, as shown in Table 5.3. This situation actually represents noise in the data that occurs as part of the data collection hardware. These kinds of outliers are also valuable to detect because the noise can be addressed to subsequently improve the accuracy of additional analysis methods. Therefore, we checked the entire Kyoto and Tulum datasets for these types of outliers. The result was that 6,398 entries from Kyoto and 9,401 entries from Tulum that represented noisy data collection conditions were removed. In the second group of experiments, all of the outliers detected by the clustering approach fit into one of these two categories. However, since we only consider the patterns which are extremely far from the cluster centroid, the rate of false negatives may be somewhat higher, which means that some real outliers are likely to be ignored by this approach. One possible solution is to decrease the pre-defined threshold, which makes our approach to detect more outliers with the risk of increasing the rate of false positives.

For the second experiment, a sequence of power readings with injected anomalies are generated by our simulation tool. To measure performance, we utilize and compare

four other existing algorithms that are used to detect anomalies in energy data. These methods are described below.

The first method, KNN density estimation (KDE), [Bellala et al., 2011] first translates time series data to frequency spectrum data and then uses a K-Nearest Neighbor (KNN) algorithm to identify anomalies in sparse regions of the data. This algorithm includes five steps. The first step is to estimate the missing values in the existing power data, which could have been caused by hardware or software failure; the second step is to compute the frequency spectrum from the imputed energy data; Step 3 calculates the distance between the frequency spectrum of power profiles between any two time windows using Euclidean distance; and in step 4, the MDS (Multi-dimensional scaling) algorithm is used to obtain a low-dimensional Euclidean of the observations from a high-dimensional space. The final step is to calculate the probability of an observation being an anomaly by applying a k-NN density estimation algorithm. The main idea of this method is that those energy points which appear in sparse regions are more likely to be identified as an anomaly. Since each observation is required to compute the distance with other observations, the time performance will be $O(n^2)$, where n is the total length of the energy sequence. The user needs to define the threshold for determining whether the observation is an anomaly.

The second method, KNN Discrete Time Warping (KDTW) [Jakkula and Cook, 2010], directly applies the K-Nearest Neighbor (KNN) algorithm to distinguish anoma-

lous data from regular energy data points. The KNN algorithm calculates the average distance for the k nearest neighbors given a specific observation. In this case, dynamic time warping (DTW) [Berndt and Clifford, 1994] is employed to measure the distance, or dissimilarity, between energy profiles. The benefit of the DTW algorithms is that it can determine the optimal match between two time series which may vary in time or speed, and hence improve the result. But the time performance of this measurement is non-linear $O(nm^2)$ to determine the distance between two sequences, where m is the length of the time window. This method is also require to the threshold to identify the anomalies in energy sequences.

Rosneer [Rosner, 1983] first proposed the Generalized extreme studentized deviate (GESD) as a potential outlier identification method. Seem [Seem, 2007] applies the GESD algorithm to examine the outliers in the energy data. This method first finds the extreme element $x_i$ which is furthest from the mean of the set. The extreme studentized deviate is determined from

$$R_i = \frac{|x_i - \overline{x}|}{s} \tag{5.6}$$

Where $R_i$ is a normalized measure of how far the extreme element $x_i$ is from the mean value $\overline{x}$ , and $s$ is the standard deviation of the elements. Next, the critical value is

determined by the equation as follows:

$$\lambda_i = \frac{(n-i)t_{n-i-1,p}}{\sqrt{(n-i+1)(n-i-1+t^2_{n-i-1,p})}} \tag{5.7}$$

Where $t_{n-i-1,p}$ is the Students t-distribution with $n-i-1$ degrees and the tail area probability $\frac{\alpha}{2(n-i+1)}$ .The number of detected anomalies is determined by the largest $i$, where the extreme studentized deviate $R_i$ is greater than the critical value $\lambda_i$.

The modified z-score (mzscore) [66] is applied to measure how far an outlier is from the mean value of the set. The typical z-score of a data point $x_i$ is

$$z_i = \frac{x_i - \overline{x}}{x} \tag{5.8}$$

where $\overline{x}$ is the mean of the set and s is the standard deviation of the set. The z-score can identify possible outliers as usual; however, this method may not be sensitive for small sizes of sample data. Thus, the modified z-score is provided as follows:

$$M_i = \frac{0.675(x_i - \overline{x})}{median(|x_i - \overline{x}|)} \tag{5.9}$$

where $\overline{x}$ is the mean of the dataset. Usually, if the value of the modified z-score is greater than 3.5, this observation can be considered as a potential outlier.

To compare these alternative methods, two versions of our pattern-based algorithm are employed. The difference between these two types lies in how we determine the threshold for identifying the outliers. As we discussed in the previous section,

the first method (PWN) defines the threshold based on a normal distribution. The threshold for the second method (PWT) is defined manually. It should be noted that the KDE and KDTW algorithms also need to define a threshold for finding anomalies. For these methods, the levels of the threshold have a large impact on the results. For comparisons sake, the optimum thresholds will be assigned for three machine-learning methods (PWT, KDE and KDTW). To attain the optimum thresholds, we manually examine the results of the methods and identify the optimum value for the best performance. From low to high, seven anomaly levels are randomly injected to the simulated energy data. The unit of the anomalies is wattage; the levels of the anomalies are from 5000 wattage to 8000 wattage. Each level include 15 anomalies. Two metrics are used to measure the performance:

(1) positive predictive value (PPV):

$$ppv = \frac{number\,of\,true\,positives}{number\,of\,true\,positives + number\,of\,false\,positives} \tag{5.10}$$

(2) accuracy:

$$accuracy = \frac{number\,of\,true\,positives + number\,of\,true\,negatives}{number\,of\,(true\,positives + false\,positives + false\,negatives + true\,negatives)} \tag{5.11}$$

Figures 5.16 and 5.17 show the results of our comparison. Overall, all five

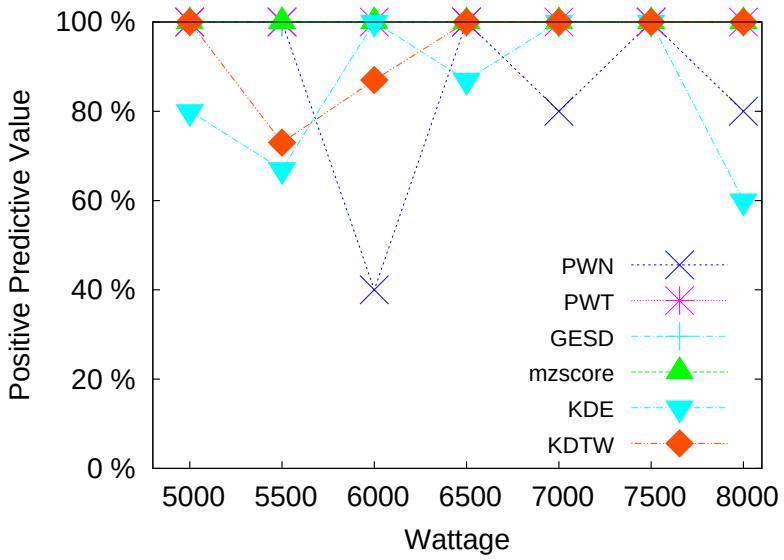



**Figure 5.16:** Comparison of positive predictive value with other methods.

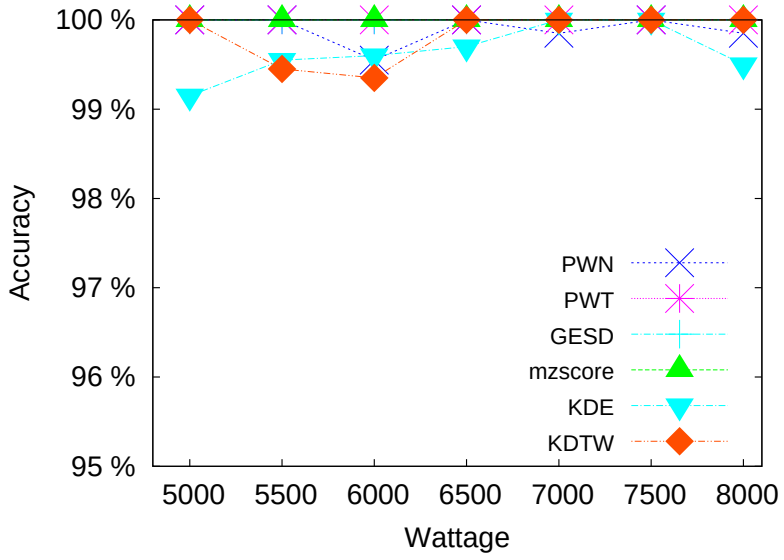



**Figure 5.17:** Comparison of accuracy with other methods.

methods have good performance both in terms of accuracy and ppv. The two statistical methods and the pattern-based method with an optimal threshold exhibit the strongest performance. All of the anomalies have been identified without false positives and negatives by these three methods. Comparing learning-based methods, the pattern-based approach with an optimal threshold achieves the best results of the four methods. Based on these experimental results, we observe that the statistical methods are still practical for detecting anomalies with high fluctuation. The reason why the performance of the learning-based methods is slightly lower for detecting extreme high outliers is that these methods identify the outliers by relying on the context of the outliers, which may mitigate the effect of the outliers.

# CHAPTER 6. CITY-WIDE BUILDING ENERGY ANALYSIS

In smart home area, our research focuses on exploring the relationship between individual home energy usage and residential behaviors. However, in the city-wide sustainable research, there are still some open questions need to be solved. How can we handle the outliers in raw energy profile? Is there any other potential features that may influence building energy usage? Or is there a way to identify similar energy profiles from a large collection of building energy profile? All these three questions will be explored deeply in this chapter. The dataset we use is a collection of building energy usage over one year in Pullman, Washington, located in the Pacific Northwest. This is part of the Smart Grid Demonstration project funded by the Department of Energy and led by Avista, a utility provider in eastern Washington state [Avista Company, 2013]. Over 15,000 Smart Meters [Avista Smart Meter, 2013] have been installed in the house for each customer as part of this project. These smart meters are deployed in an Advanced Metering Infrastructure (AMI), which makes electricity readings available every five minutes. The data consists of a customer account number for each building and a set of electricity readings for the building with the corresponding date and time at which the reading was observed. Additional information

is provided for each meter site including latitude and longitude and the type of building (residential, industrial, or commercial) that is located at the site. The purpose of this work is to use machine-learning and data-mining techniques to discover patterns, cycles, and trends of energy consumption for large-scale deployments, and also to further explore potential building features that correlate with energy consumption.

## 6.1 Data Format

Before explaining the algorithms, data types and formats are introduced for further analysis. These data are collected from building smart meters every five minutes. The power values (kWh) per building are automatically sent to a remote database at Avista. Table 6.1 illustrates the basic format of smart meter readings, including the smart meter identifier, the time the events are collected, and the electricity that was consumed over the last five minutes. In addition, geographical information for the smart meter represented by the corresponding latitude and longitude is provided as shown in Tables 6.1-6.3. It can be seen that four different types of buildings are listed: (1) Residential: normal residential houses and apartments; (2) Industrial: Washington State University buildings as well as a limited number of independent manufacturing sites; (3) Commercial: all kinds of service industry buildings such as stores, restaurants, and banks. (4) Mixed: the specific buildings including both res-

idential and commercial purpose. Moreover, the address of each building and the corresponding smart meter are provided, as shown in Table 6.2. By importing the postal address, the basic building information can be attained from public tax assessors, including the building value, size, age, materials, and geographic region. It should be noted that the address listed in Table 6.3 is not a real address. We changed the address in this table to maintain the privacy of the customer.

**Table 6.1:** Example of a power event.

| **SmartMeter ID** | **DateTime** | **Value(kWh)** |
|---|---|---|
| 13988720 | 2011-04-21 08:00:00 | 0.032 |

**Table 6.2:** Geography and type of smart meters.

| **SmartMeter ID** | **Latitude** | **Longitude** | **Type** |
|---|---|---|---|
| 13986534 | 46.7391 | -117.176 | Residential |
| 13993727 | 46.7956 | -117.177 | Industrial |
| 13987294 | 46.7375 | -117.175 | Commercial |
| 500002312 | 46.7341 | -117.168 | Mixed |

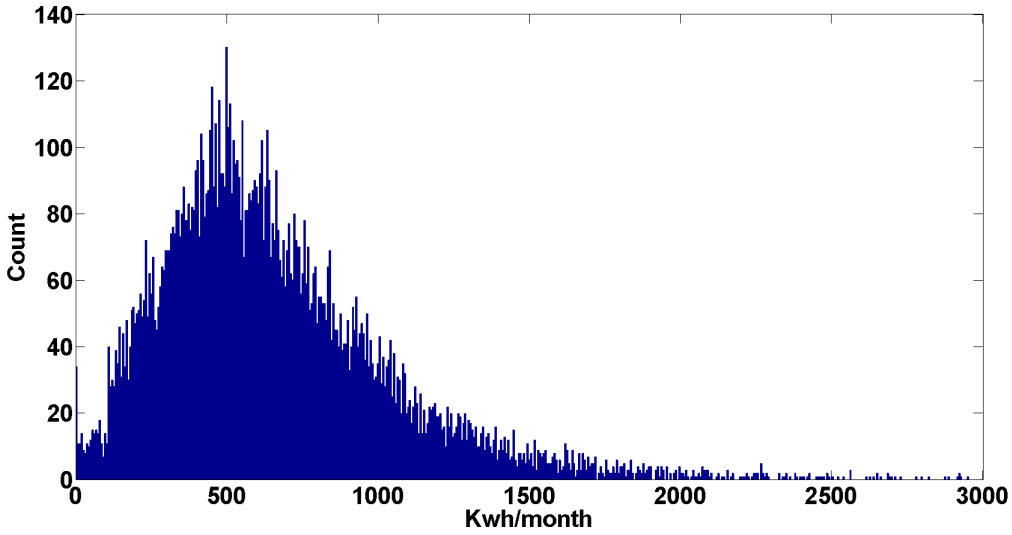
A preliminary analysis of energy data is used to explore the distribution of building energy usage. Figure 6.1 shows the histograms of the total energy consumption per month for all the buildings in our data set. The difference between the plots

**Table 6.3:** Postal address of a building and its corresponding smart meters.

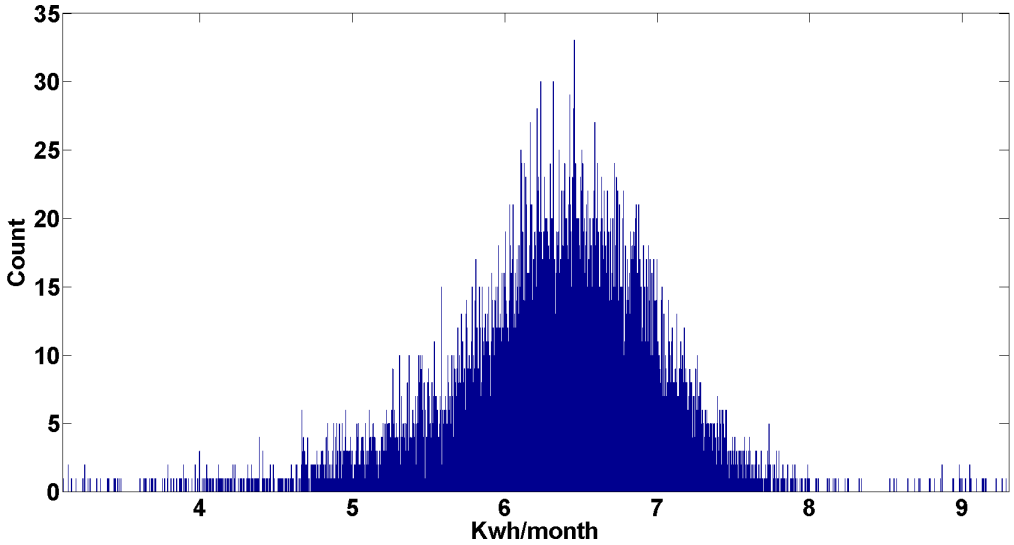
| SmartMeter ID | Address | Postal | City |
|---|---|---|---|
| 13986513 | 2100 Pullman Road | 99163 | Pullman |

is that the x axis of Figure 6.1(a) is on a regular scale, and the axis of Figure 6.1(b) is on a logarithmic scale. Figure 6.1(a) illustrates that the energy usage per month for all the buildings roughly follows a log-normal distribution skewed to the left. That means that, for most of the buildings, energy usage keeps at a relatively low level, but some buildings in the high percent of energy usage consume several times as much energy as the buildings in the low percent. As shown in 6.1(b), energy usage on the logarithmic scale follows the Gaussian distribution much more closely than in Figure 6.1(a).

In fact, there are various branches of phenomena in economics and science that roughly follow log-normal distributions [Newman, 2005], including some features that may influence energy consumption, such as income and age of marriage. In particular, researchers [Kolter and Jr., 2011] pointed that building energy usage also roughly follows a log-normal distribution. In addition, power law theory defines a potential linear relationship between the logarithms of input and output variables. This type of relationship exists in many areas, such as the size of cities and population size.




(a) Regular scale




(b) Logarithmic scale

**Figure 6.1:** A histogram of total energy consumption per building on the x axis.

The location-based smart meter data is difficult to be understood directly by reading raw values. To represent these data, a web-based visualization tool is presented in the next section.

## 6.2 CASAS web-based city-level Energy Visualization

The smart meters that we monitored generate extremely large volumes of data every day. It is not practical for researchers and analysts to extract useful information by examining such large amounts of raw data. A city-level energy visualization tool, EnergyViz, is designed for monitoring and visualizing real-time energy usage generated by location-based smart meter data installed in the buildings. The system architecture of EnergyViz is shown in Figure 6.2. The database stores all of the needed information for the visualization. The web server is responsible for retrieving meter readings and geospatial data from the main database and providing the web-based client with suitable information to display energy usage for a specific geo-location based on different dimensions such as time, location, and level of energy usage.

EnergyViz's client side integrates with the Google Maps API [Google Map, 2013] for geospatial energy usage representation. A screenshot of our web-based city-wide energy usage visualization tool is shown in Figure 6.3. EnergyViz colors each smart

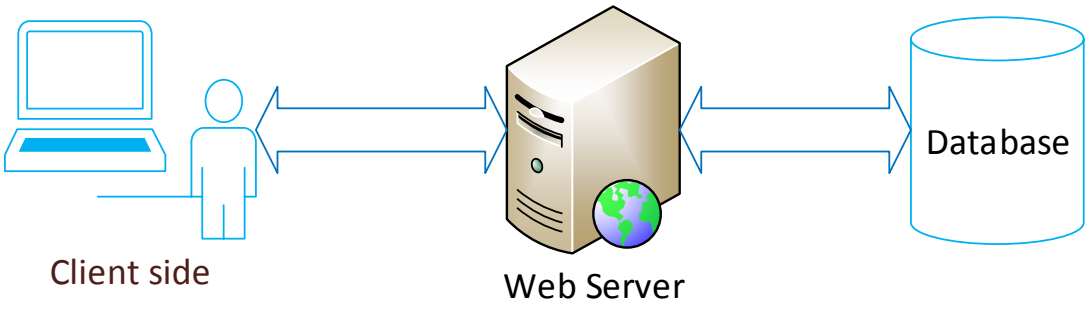



**Figure 6.2:** System architecture of EnergyViz.

meter location according to its usage over the collection period (hotter colors such as yellow and red indicate greater energy usage). Moreover, the tool also can mark the location of the neighbors for a specific user. That information will be helpful to allow users to compare their energy usage with their neighbors as shown in Figure 6.4.

## 6.3 Outlier Detection

In real-world data collections, one of the most common problems is that the outliers existing in raw data may interfere with the analysis of energy usage. Figure 6.5 indicates a time-series chart of power usage collected from a single building. From the figure, we see that the values of the meter readings stayed at a very small value most of time. However, some of the values increased abruptly, which are much larger than the vast majority of the readings. These outliers can cause inaccurate results for

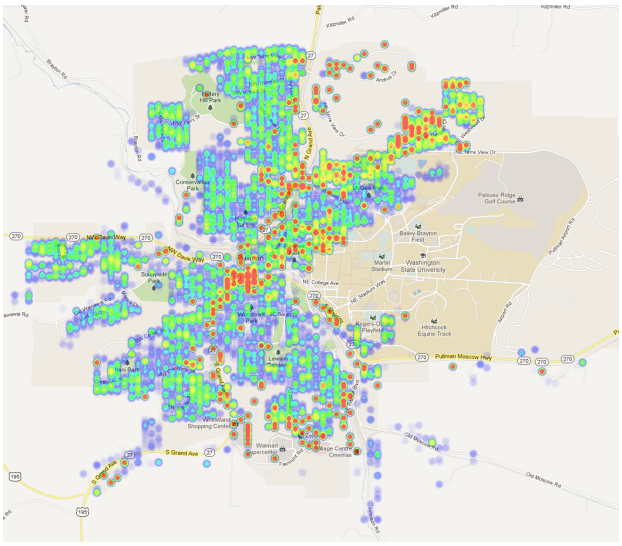


**Figure 6.3:** Image of the city-level energy visualization tool.

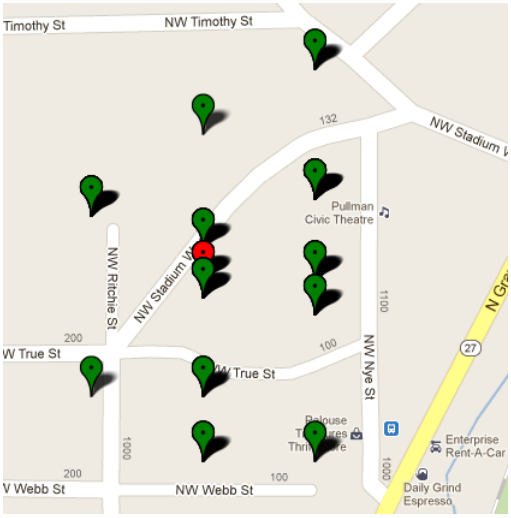



**Figure 6.4:** Image of the building (red marker) and its neighbours (green marker).

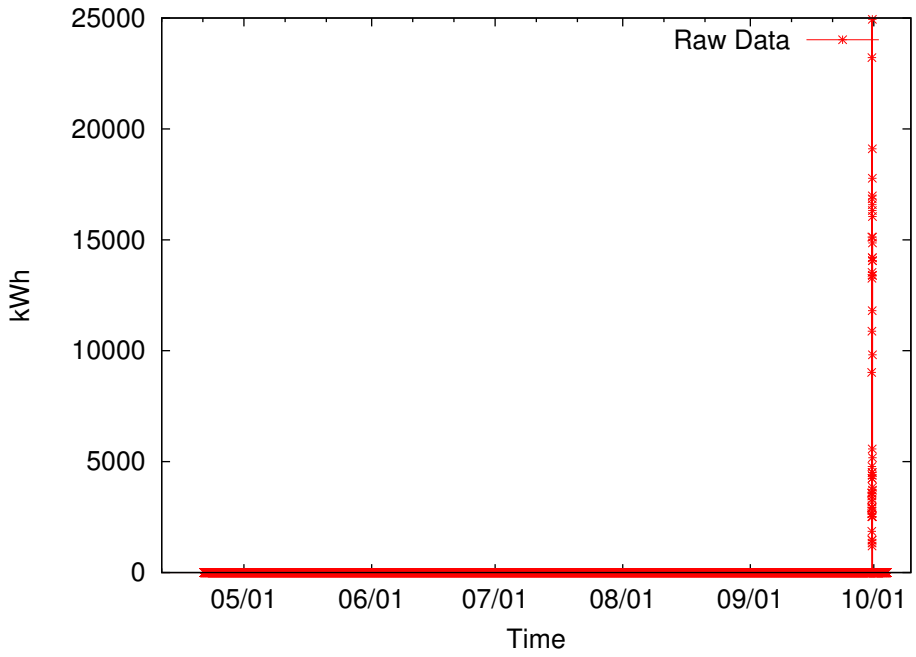



**Figure 6.5:** Extreme outliers in raw data collected from a building.

machine learning algorithms. Avista verified that these readings are in fact indicative of noise and should not be considered as valid readings. To remove these outliers, we use regression models to predict the likely actual values that occurred during the times that noisy readings were generated.

The main components of smart meter readings are date, time, and energy usage. The potential relationship between time and energy usage needs to be explored by regression models. To predict energy usage, the time-based features are considered as the input of the models. The time-based features are listed in Table 6.4.

**Table 6.4:** List of time-based features for outlier detection.

| **Time-Based Feature** |
|---|
| Month |
| Date |
| Day Of Week |
| Hour |
| Minute |
| Time Of Day |

The output of the algorithms is the predicted amount of electricity that will consume at the indicated date and time. Machine learning algorithms are capable of learning and recognizing complex patterns based on the input features. In this work, regression models were used to map these time-based features onto the amount of energy the residents may consume. The regression model can produce the continuous-valued output, which allows users to understand the result more clearly. While the predicted activity may not indicate exactly the amount of energy consumed, the purpose of this work is to generate a reasonable value to replace the outliers in order to remove erroneous readings and improve the overall performance of the analysis algorithms.

### *6.3.1 Regression Models and Experimental Results*

Since there are more than 15,000 energy profiles in the smart meter database, it is impractical to test the algorithms on all these profiles. In this case, 10 different energy profiles are randomly selected for our experiment. Four regression models are chosen to predict energy values: (1) linear regression; (2) a support vector machine with a linear kernel; (3) a support vector machine with a RBF-kernel; (4) a support vector machine with a polynomial kernel. These methods have already been introduced in Section 4; thus, it is not necessary to discuss the details of these algorithms in this section. To evaluate the performance of the algorithms, 10-fold cross validation was applied. The algorithms were evaluated using two metrics: (1) correlation coefficient, which measures how a regression model fits the data sets and (2) root mean squared error (RMSE), which measures the difference between the values predicted by the model and the actual values.

Figure 6.6 and 6.7 show the performance of alternative algorithms on 10 different energy profiles. The performance of the linear regression algorithm and the SVM with linear and RBF kernels are very close both on correlation coefficient and RMSE. The SVM with RBF kernel is only marginally better than the other two. All of these algorithms perform much better than the SVM with non-linear model. We argue that the linear models and RBF kernel are preferable for energy prediction.

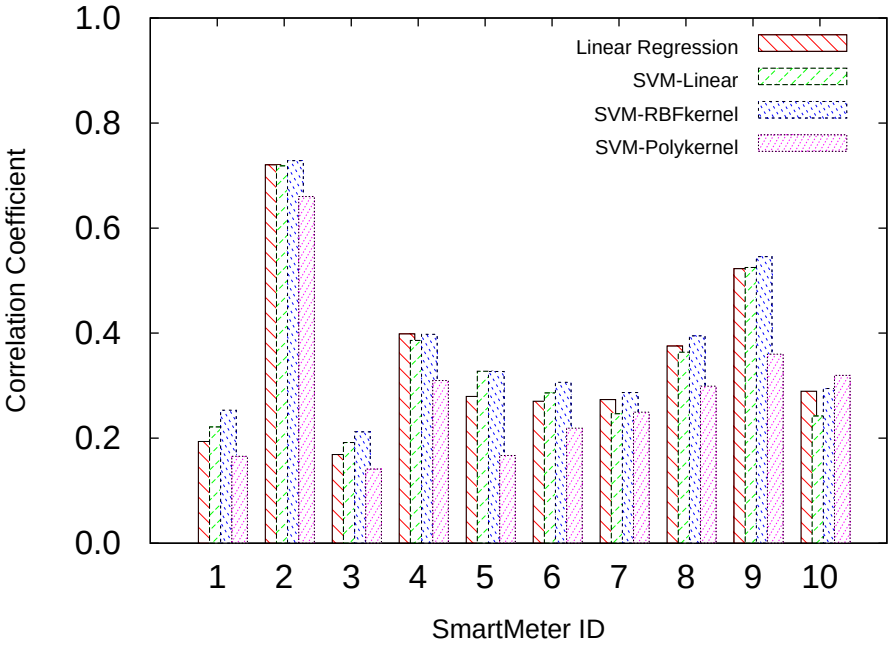



**Figure 6.6:** Correlation efficient of the performance of different algorithms and energy profiles.

Comparing with the performance between the energy profiles, the performances vary greatly depending on the various energy profiles. Some of them fit the regression models very well, and some of them are really difficult to model.

Figure 6.8 shows a clean time-series power values coming from the same building shown in Figure 6.5 after removing the outliers. It can be seen that the data has been mitigated and is much smoother. Note that while the y axis in Figure 6.4 ranged from 0 to 25,000, the y axis in Figure 6.8 ranges only from 0 to 0.8, which is much

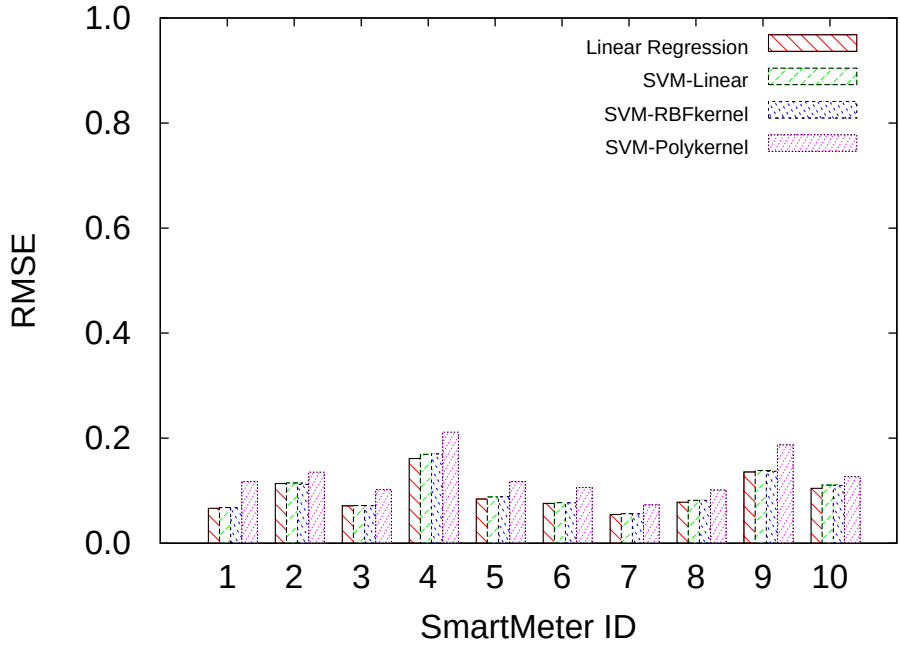



**Figure 6.7:** RMSE of the performance of different algorithms and energy profiles.

more reflective of actual usage values. Based on the linear regression algorithms that we tested, an outlier detection tool was developed to detect and fix the outliers automatically. Through scanning the whole dataset, a total of 124,206 outliers were detected and fixed by our tool. This work lays the groundwork to develop related algorithms and techniques for future work.

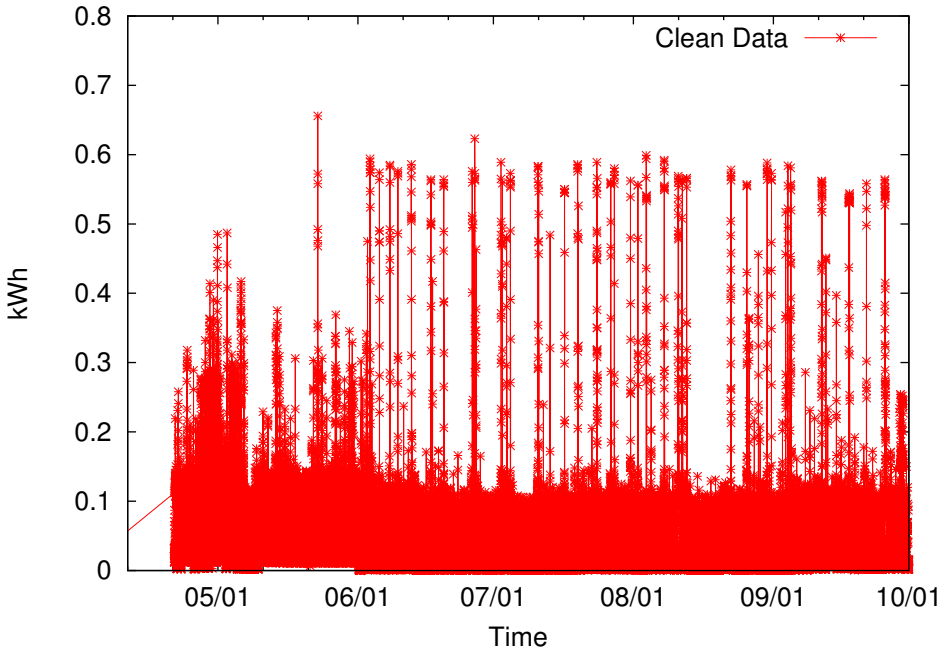



**Figure 6.8:** Plot of clean data after removing the outliers.

## 6.4 Energy Prediction Using Building Features

In this section, we use a data-driven approach to model energy usage of residential buildings. The input of our models are various building features that were collected from the publicly-available tax assessor records [TerraScan TaxSifter, 2013]. We hypothesize that providing users with knowledge about the relationship between building features and energy consumption and giving them suggestions for energy reduction will result in more substantial decreases in overall consumption. In par-

ticular, we use machine learning methods to correlate building features with energy usage and provide important feedback to promote energy-efficiency building design. First, we develop a tool to automatically collect building features from the public tax assessor web site and to correlate energy usage with these features. In addition, we analyze the distribution of residential home energy consumption, and present both linear and non-linear regression models over the datasets on the regular and logarithmic scale separately. Furthermore, we compute the coefficient correlation between the individual features and energy usage to determine the most relevant features. Finally, we use a web-based application to provide estimated energy usage given the building features, and compare usage with other similar buildings.

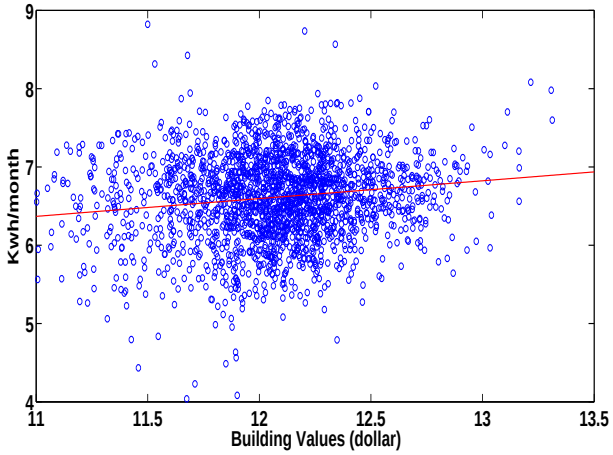
We develop an automatic tool to attain building records from a publicly-available tax assessor web site. The tax assessor records contain detailed information about registered properties and buildings in Pullman and a list of building features for each site, such as the values and areas of the building, the year the building was built, building quality and condition, and other similar features. There are some special situations that need to be handled separately. For example, in the case where multiple apartment units are located at the same street address. it is difficult to make a distinction between their energy usage. Thus, in our experiment, we only select readings that occur at a unique street address. After using addresses to link energy usage and tax assessor-based building features, there are a total of 2,090 unique buildings

that we include in our data set.

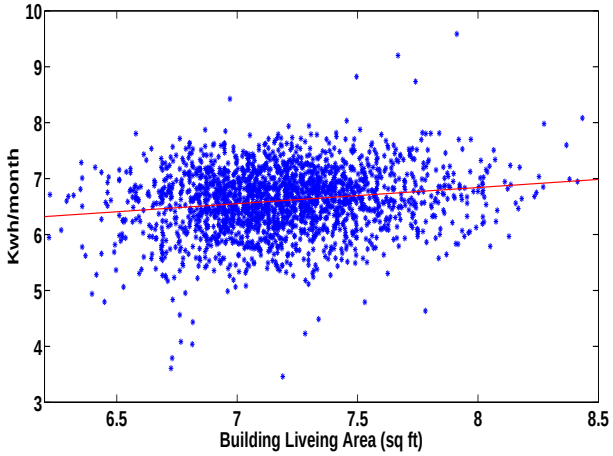
We assume that there may exist a similar relationship between building features and residential usage. Figure 6.9 shows the relationship for our data set, plotting building values and living area separately on a log-log scale. Although there exists plenty of noise in real data, there is basically a linear relationship between input and output, indicating that a potential power-law relationship exists in the data set. The next section will further exploit different building features that are used to derive predictive models of energy consumption.

### *6.4.1 Data Features*

Before illustrating our energy prediction techniques, we summarize the 73 total features we extract from the tax assessor site. All of the features are real-valued and we include them directly in the final feature vector. To train our algorithms, we can simply put all the features into our models for predicting energy consumption. Nevertheless, for the sake of the customer, it is very practical to determine which of the features are most relevant to energy consumption. For example, knowledge of which materials correlate with lesser energy usage may be helpful to residents to reduce their energy usage. To illustrate this, Table 6.5 and 6.6 shows the computed correlation coefficient between each of the different building features and monthly




(a) Building values




(b) Building living area

**Figure 6.9:** Plots of building values and living area versus monthly energy consumption, both in logarithmic scales. (Building Values (a), Pearsons correlation coefficient: 0.192; below: Building Living Area (b), Pearsons correlation coefficient: 0.165) The red line shows the least-squares fit.

**Table 6.5:** Positive correlation coefficient between energy usage and building features in isolation.

| **Building Features** | **Correlation Coefficient** |
|---|---|
| Number of the units | 0.3801 |
| Land Value | 0.3687 |
| Living Area | 0.3671 |
| Ceramic Tile | 0.3487 |
| Total Building Value | 0.3023 |
| Improvement Value | 0.2753 |
| Metal Preformed | 0.2352 |
| Masonry Stucco on Block | 0.1719 |
| Floor Insulation Area | 0.1489 |
| Quality (level 1- level 6) | 0.1182 |

energy consumption. It should be noted that the only top 20 features are listed in the table. The list of all building features is provided in Appendix A.

These values capture how each feature is related to energy usage independent of other features. The features above the double lines are positively correlated with energy consumption and the below are negatively correlated with energy usage. Both

**Table 6.6:** Negative correlation coefficient between energy usage and building features in isolation.

| **Building Features** | **Correlation Coefficient** |
|---|---|
| Hardwood | -0.0538 |
| Forced-Air Furnace | -0.0443 |
| Composition Shingle | -0.0427 |
| Frame Siding Wood | -0.04 |
| Number of Bathrooms | -0.039 |
| Vinyl Sheet | -0.0354 |
| Attached Garage | -0.0336 |
| Life of the buildings | -0.0278 |
| Frame Wood Shingle | -0.0259 |
| Number of Bedrooms | -0.0251 |

of these sets of features can be inferred to have an impact (positive or negative) on energy expenditure. It makes sense that the building living area, building value (this includes land value and improvement value), and quality have a strong positive relationship with energy usage. Some specific materials, such as ceramic tile, metal preformed, and masonry stucco also greatly influence energy consumption. However, it is not surprising that floor insulation has a positive influence on energy usage. The reasonable explanation is that houses with more floor insulation may be also large, and thus consume more energy consumption on average. In regarding to negatively correlated features, the life of the buildings, number of bathrooms, bedrooms, and garage are main features decreasing the energy usage. That means the newer houses and more rooms will cost slightly less energy. Materials such as hardwood, furnace, composition/wood shingle, and frame siding wood, can be also useful to mitigate energy usage.

In addition, to measure the strength of a power-law relationship in the data, we did another analysis of building features and energy usage on a logarithmic scale based upon the discussion in Section 6.1. It should be mentioned that we aggregate some sets of building features into a single type to avoid sparse features. For example, multiple exclusive frame material binary features are aggregated into one multi-valued frame material feature. After combining the features, the new set of building features will be converted into a total of 34 features.

### *6.4.2 Predictive Models*

We hypothesize that machine learning techniques can be used to predict energy consumption based on information about building features. Here we describe the methodology we use to validate this hypothesis. The input and class feature values for the machine learning algorithms consists of building characteristics and energy usage, provided both on a linear scale and (in keeping with the power law relationship discovered in the data) on a logarithmic scale. The purpose of the learned models is to provide users with basic information about their energy consumption based upon characteristics of their building. Three regression models are applied to predict energy usage. Details of the regression models are introduced in Section 4.2.3. The linear regression model and the support vector machine with a linear kernel are used to explore the linear relationship between more input features and energy usage. The support vector machine with a non-linear kernel is evaluated for its ability to detect energy usage in the building related to known features in the nonlinear relationship.

### *6.4.3 Experimental Results*

Two series of experiments were performed for energy predication. The first experiment uses building features on a linear scale. In the second experiment, the

**Table 6.7:** Cross validation performance of the different algorithms on a linear scale.

| Time | Linear Regression | | SVM (Linear) | | SVM (Non-Linear | |
|---|---|---|---|---|---|---|
| | Correlation Coefficient | No Log RMSE | Correlation Coefficient | No Log RMSE | Correlation Coefficient | No Log RMSE |
| Day | 0.412 | 36.473 | 0.3447 | 37.993 | 0.294 | 42.564 |
| Week | 0.409 | 254.420 | 0.340 | 264.440 | 0.301 | 291.160 |
| Month | 0.399 | 1093.200 | 0.331 | 1139.800 | 0.292 | 1251.240 |

same features are converted to the logarithm scale. We evaluate the performance of the algorithms using 10-fold cross validation and report average error and correlation coefficient. Tables 6.7 and 6.8 show the performance of alternative algorithms on two data sets. The algorithms were evaluated by two metrics: (1) correlation coefficient and (2) root mean squared error (RMSE) on the energy usage. Three different time periods were selected for testing the performance: day, week, and month.

In regard to the two tables, the linear regression model obtains the best overall performance both on correlation coefficient and RMSE. The SVM with a linear kernel is marginally worse than the linear regression model, which performs much better than the SVM with non-linear kernel method. We argue that the linear models are preferable for predicting energy. The results also provide evidence to validate our

**Table 6.8:** Cross validation performance of the different algorithms on a logarithmic scale (CC: Correlation Coefficien).

| Time | Linear Regression | | | SVM (Linear) | | | SVM (Non-Linear | | |
|---|---|---|---|---|---|---|---|---|---|
| | CC | Log RMSE | No Log RMSE | CC | Log RMSE | No Log RMSE | CC | Log RMSE | No Log RMSE |
| Day | 0.273 | 0.581 | 35.960 | 0.225 | 0.5918 | 29.045 | 0.143 | 0.820 | 58.272 |
| Week | 0.271 | 0.581 | 244.311 | 0.223 | 0.592 | 206.707 | 0.142 | 0.8197 | 394.621 |
| Month | 0.270 | 0.580 | 1015.800 | 0.222 | 0.592 | 863.447 | 0.140 | 0.820 | 1705.361 |

hypothesis that building features have a strong linear relationship with energy usage in buildings.

In comparing the effect of various time windows on prediction performance, the performance for a daily time window is only slightly better than the two other time windows. In comparing the overall performance between the linear and logarithm scales, energy consumption on the log-based data set has the better performance based on RMSE in the linear scale. Since the data is not on the same scale, the correlation coefficient cannot be directly comparable. This result also provides evidence that a power-law relationship does exist between building features and energy consumption.

### *6.4.4 Web-Based Energy Prediction Interface*

The last component of our work is a feedback tool to promote energy savings for users. Based on regression models described in the previous sections, an end-user application is developed as a web-based solution and can therefore run on a computer display or a mobile device. Our system will estimate monthly energy usage based on the building features that the users input. Nevertheless, it is not convenient for most users to input all these features to our learning models. From Appendix A, it can be noted that some features are not greatly relevant to energy usage. Thus, for the sake of simplicity and intuition, it is necessary to determine which subset of features most greatly influences energy usage. To identify these features, we use a feature selection method, minimum-Redundancy-Maximum-Relevance [Peng et al., 2005], to select a set of features which are far away from each other and still maintain high relevance to the final target.

Figure 6.10 shows the final list of building features we select for predicting energy usage in our tool and a bar chart for letting users view and compare predicted energy usage generated by our learning algorithms and their real energy consumption. Since comparing users own energy usage with others can promote energy-efficient behaviors, our system also provides the functions that compare their energy consumption with people living in similar buildings. Here, we use an unsupervised learning algorithm

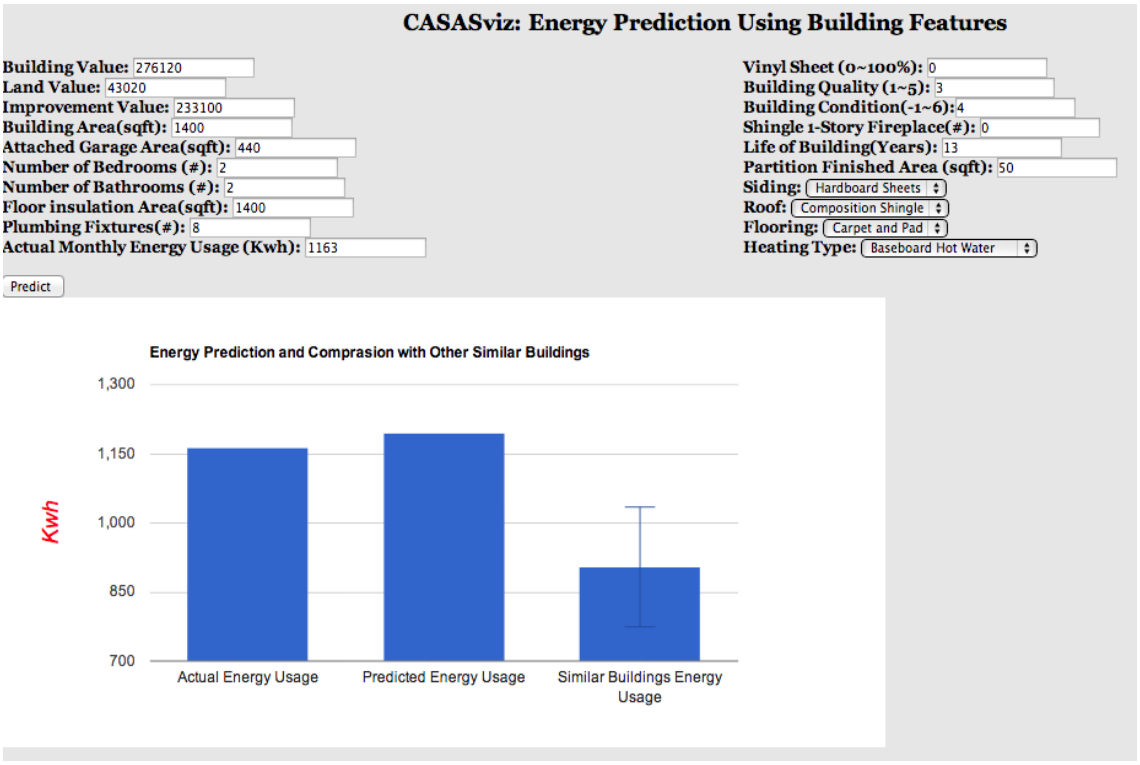



**Figure 6.10:** Web-based user interface for estimating building energy.

to separate the total buildings into several clusters. Similar buildings will be assigned into the same cluster. In Figure 6.10, the left figure shows a buildings consumption and displays where this building lies in the overall distribution of similar buildings in the same cluster. Since our approaches are fully data-driven, our models are applied particularly to the residents living in the Pullman, Washington, area. In this chapter, we consider the main influence of building features and structures on the residential energy consumption. We further use machine-learning methods to map a

significant number of features of the buildings with their corresponding energy usage. The web-based tool is allow users to input building features for attaining estimated energy usage and compare it with other similar buildings. However, how to find out the similar users from a large collections of time-series smart meter data is still a challenging problem. A segmental clustering algorithm will be presented to solve this problem in the next chapter.

# CHAPTER 7. SEGMENTAL CLUSTERING ALGORITHM

The sources of usage information of smart meter data not only reflect differences in collection methods but also reflect differences in the population. We want to determine to what extent usage patterns are common among these population subgroups and data sources, and use the similarities to make inferences. To do this, we will draw on techniques from collaborative filtering [Adomavicius and Tuzhilin, 2005, Su and Khoshgoftaar, 2009]. Collaborative filtering has been widely used in recommender systems such as Amazon [Linden et al., 2003] and Netflix [Tscher et al., 2008] to offer recommendations for one individual based on the similarity of their profile to others who liked the same product.

There are a number of measures to determine similarity of users from data sources. The most common technique is the clustering methods based on unsupervised learning algorithms. However, the efficiency of the existing clustering algorithms performs pretty low on a large scale of real-time datasets.

As shown in Figure 7.1, the two user profiles are fairly similar most of time. From a global perspective, these two user profiles could be viewed as similar consumers. However, during a small period, a large fluctuation appears in one user profile. That is very common in normal users. The users may go outside for traveling or the

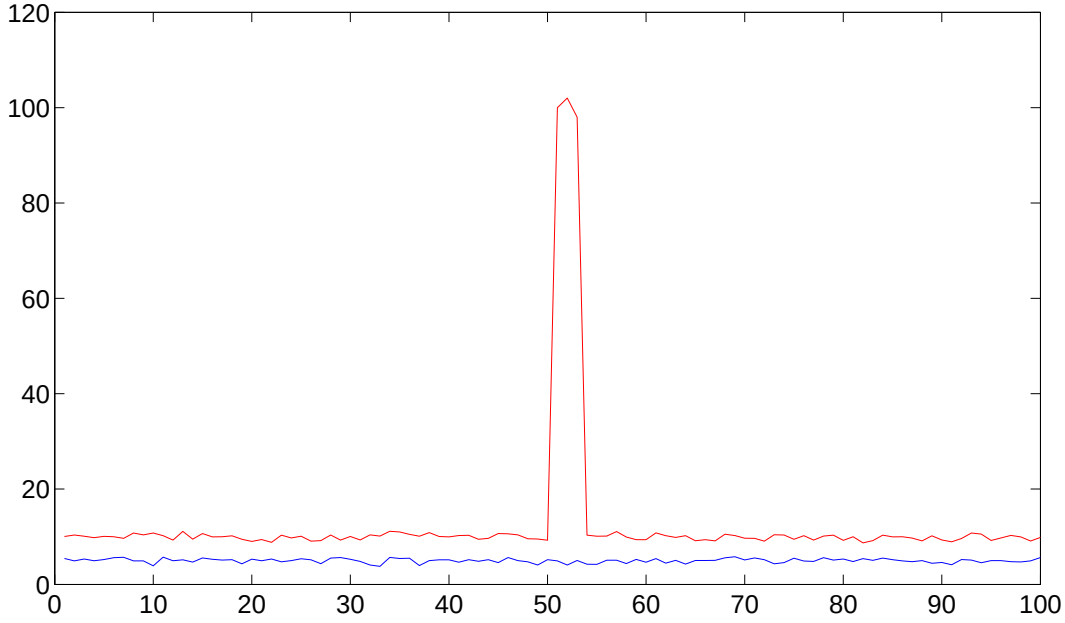



**Figure 7.1:** Example of the global dissimilarity caused by local fluctuations.

parties may be open in their houses. However, in the aspect of regular similarity measurements, these two users will be identified as non-similar users due to the overall huge similarities between their user profiles.

On the other hand, the time-series user profiles will continue to grow. Another problem with regular similarity measurement that may arise is double counting, because the previous user profiles will also be re-calculated when the new profiles show up. As shown in Figure 7.2, the blue figure represents previous user profiles; the orange one indicates new user profiles. From the figure, we can see that the similarity of two user profiles is re-calculated once new profiles are generated.

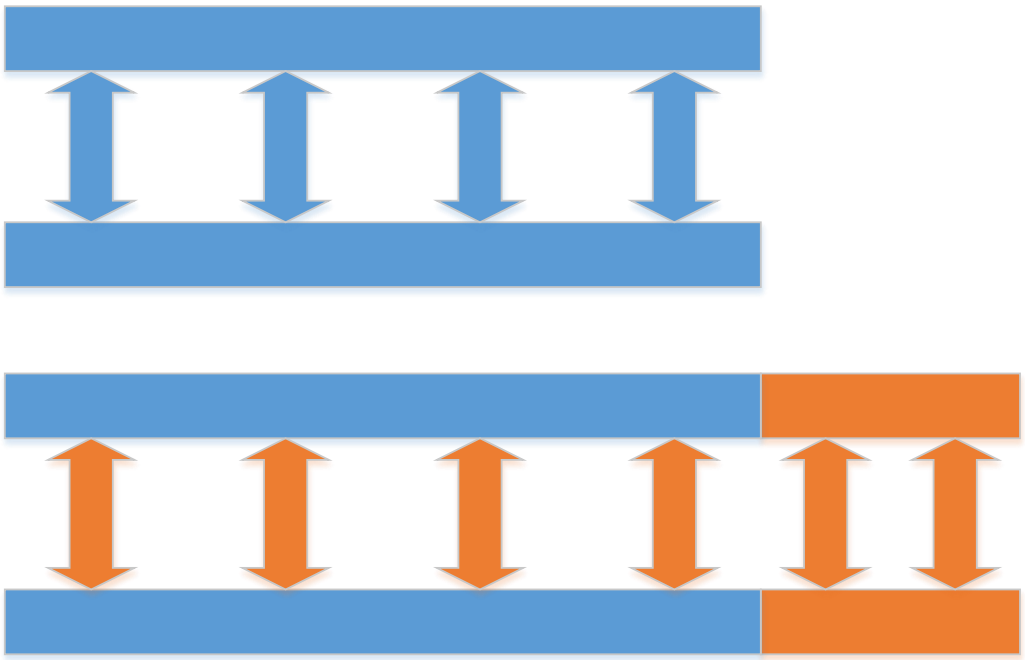


**Figure 7.2:** Demonstration of double counting caused by new profiles.

To avoid problems like local fluctuations and double-counting, a segmental clustering algorithm is designed to handle these issues. The method first splits a long sequence of user profiles into several segmental sequences. For each segment, the unsupervised learning method will be applied to cluster similar user profiles, and then a matrix of similarity weights will be generated. Similarity weights measure the similarity between two users: a large weight indicates that the users are close. In the next step, only a matrix of similarity weights is required to update the values without re-calculating the previous weight. Figure 7.3 shows how our method clusters user

segmental profiles based on various periods.

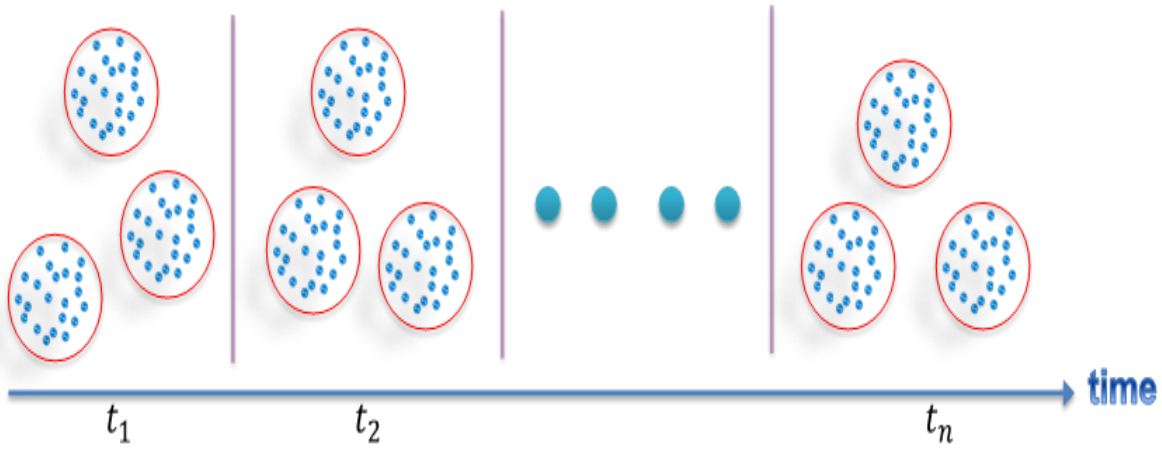



**Figure 7.3:** Demonstration of segmental clustering algorithms.

## 7.1 Methods

Power data is measured by a single meter as a single power-time sequence. In our smart meter datasets, each smart meter will generate a power-time sequence respectively. In our algorithms, a long-period power-time sequence will be divided into equal-size sub sequence during a fix period (i.e. day, week, and month). Euclidean distance is applied to measure the dissimilarity of the frequency spectrum of the two power-time sequences. Our proposed method consists of three steps:

**(1) Transform Time Domain to Frequency Domain**

The Fourier transform [Bracewell, 2000] is a common method to transform data from a function of time, $f(t)$, into the frequency spectrum, $\widehat{f}(\varepsilon)$. The strength of the Fourier transform is that it can divide complex time signals into several frequency signals on various phases. The method allows complex signals to be compared more accurately and conveniently. In our method, the fast Fourier transform (FFT) [Cooley and Tukey, 1965] is applied to generate the frequency spectrum. Given a power-time sequence, $x[n]$, for $n = 1, , N$, the frequency spectrum can be calculated as

$$X(K) = \sum_{i=1}^{n} x(n) e^{-i2\pi k \frac{n}{N}, k=0,\ldots,N-1} \tag{7.1}$$

**(2) Cluster Power-Time Sequence**

In order to compare $M$ different power-time sequences generated by $M$ different smart meters during a fixed period, we use an unsupervised clustering algorithm. In particular, we utilize a k-means algorithm [Hartigan, 1975] to cluster user frequency spectrums. In this case, $k$ indicates the number of clusters in the algorithm. One of the most challenging problems in k-means is to decide upon the best $k$ to yield optimal clusters. The choice of $k$ is often hard to decide, highly depending on the distribution of data points in the data set. In our method, we first estimate the umber $k \approx \sqrt{n/2}$ based on a Kantis recommendation [Mardia et al., 1980], then use the Silhouette method [Rousseeuw, 1987] to provide the choice of $k$. For a given

cluster, the value of Silhouette $S(i)$ is defined as follows:

$$s(i) = \frac{(b(i) - a(i))}{Max(a(i), b(i))} \tag{7.2}$$

Where $a(i)$ is the average dissimilarity between the point $i$ and all other points within the same cluster; $b(i)$ is the minimum average dissimilarity between the point $i$ and other data of other clusters. The $s(i)$ approaches a value of 1, the data point because more appropriately clustered. The average $s(i)$ indicates a measure whether all the data has already been clustered. For each period, we use the Silhouette method to select one local optimum $k$ in a given range $(\overline{k} + \sigma, \overline{k} - \sigma)$. The $\sigma$ is a parameter that the user specifies. The global optimum $k$ can be attained by computing the average of all local optimum $k$ in the whole period.

**(3) Calculate and update the weight**

To measure the long-term similarity of two users, two parameters are designed. The first parameter is called the similarity weight. This weight can be considered as a measure of the similarity between two user profiles within the same cluster, which is defined as:

$$w_{ij} = \frac{d_{max} - d_{ij}}{d_{max}} \tag{7.3}$$

Where $d_{ij}$ indicates the distance between the users $u_i$ and $u_j$ and $d_{max}$ is the maximum of all the distances between the users. Here a larger weight means closer users. For

users belonging to different clusters, the weight will be set to zero. In the next period, a new weight $w_{ij}^{'}$ will be generated. The total weight will be updated as $w_{ij} + w_{ij}^{'}$.

The second parameter is a measure of similarity frequency $f_{ij}$ of two users $u_i$ and $u_j$ being in the same cluster. If two users are assigned into the same cluster, the frequency $f_{ij}$ will be upgraded to $f_{ij} + 1$. The benefit to use these two parameters is that it does not need to repeat the process of computing users similarity as done in earlier periods and reduces the influence of local fluctuation due to unexpected causes.

## 7.2 Experimental Results

The dataset we use to do the experiments is generated from Avista smart meters over 49 weeks. Each smart meter collects the kilowatt hour power every five minutes. The raw datasets are extremely large, which cannot be processed and analyzed by regular personal computers. To utilize the dataset more efficiently, the five-minute power readings are aggregated to daily power values. In addition, some of smart meter data is incomplete over 49 weeks due to inconsistent installation time. Those incomplete data are excluded in the experiments. The final dataset includes 11,878 different residential smart meter data points over 49 weeks, and each week is regarded as the segmental period for the experiment. It should be noted that ground

truth is lacking for labeling similar user profiles. To validate the performance of our algorithms, three pairs of user profiles are selected. The first pair of user profiles maintains the close relationship over the whole time; the second two user profiles keep apart over the whole time; in the last pair, two user profiles keep similar over 30 weeks and separate during the remaining time.

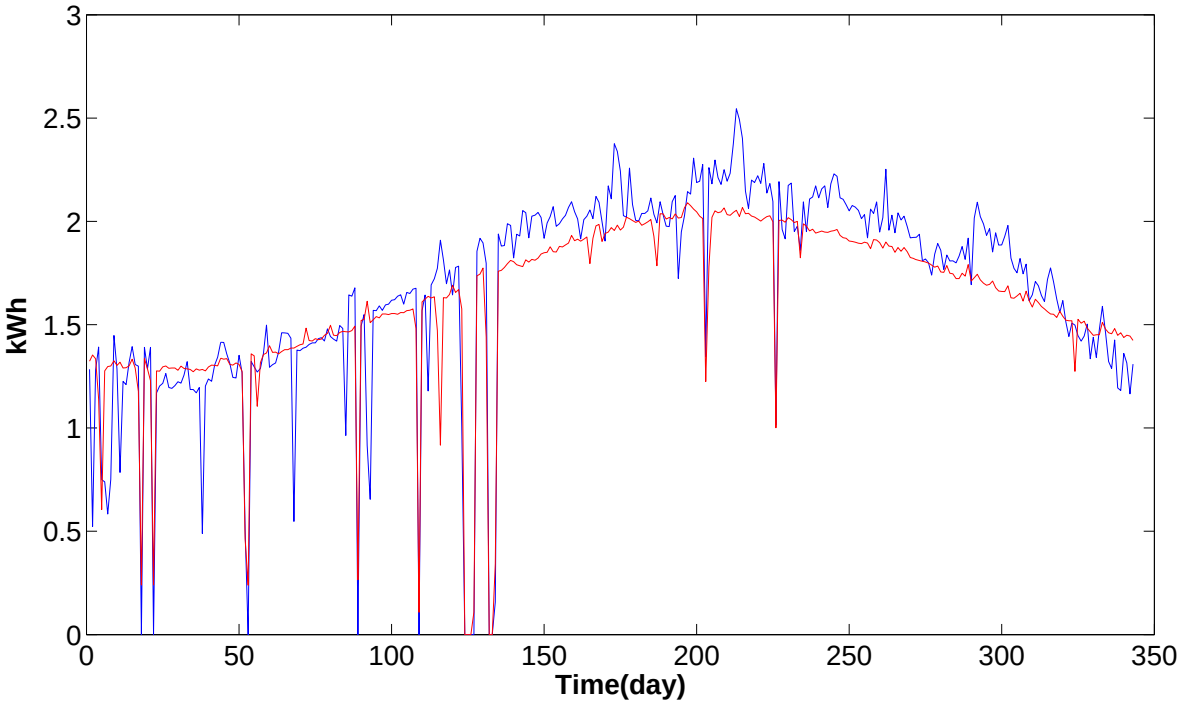



**Figure 7.4:** Two user profiles are within the same cluster over 49 weeks.

The similarity frequency $f_{ij}$ of the first two user profiles are 49. That means these two user profiles are assigned into the same cluster over the whole 49 weeks. Figure 7.4 shows the time-series line chart of two user profiles. From the figure, we

see that these two user profiles maintain close in most of time. These results indicate that the similarity frequency can identify two user profiles with long-term similar relationships.

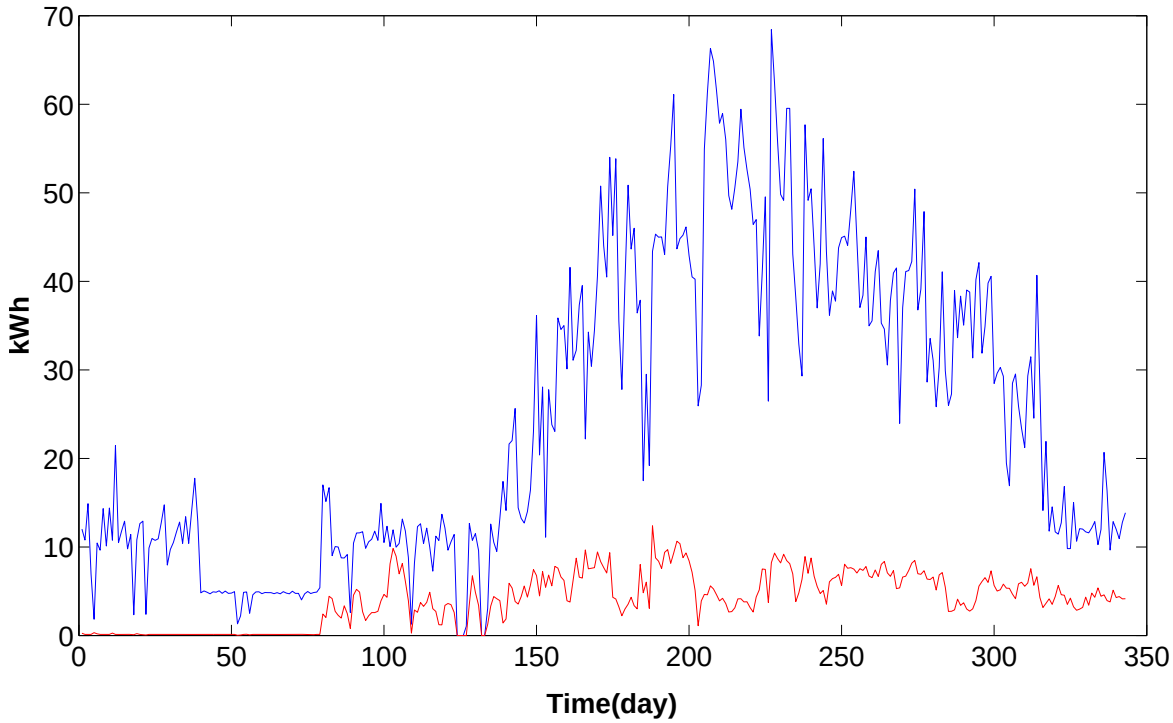
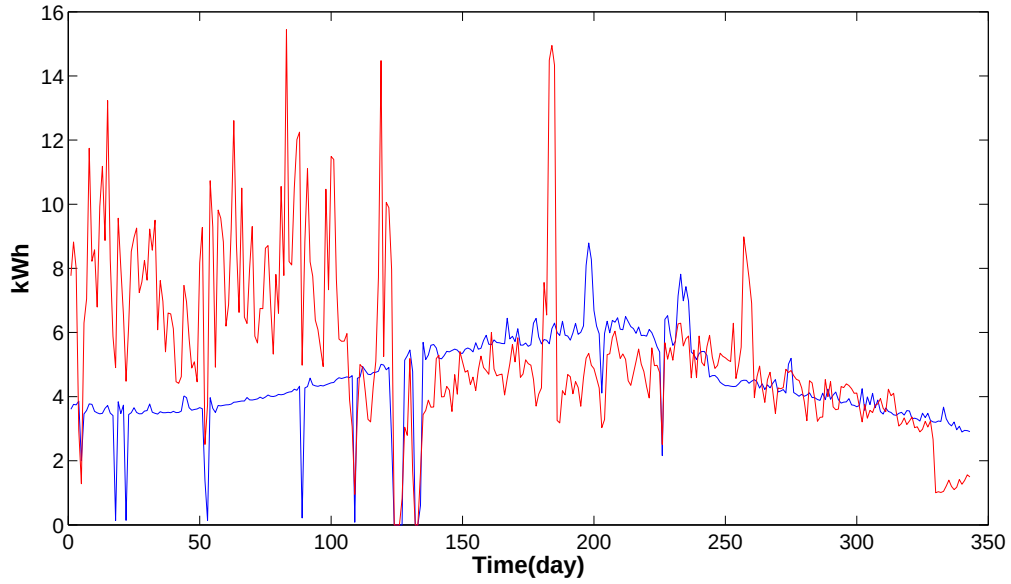
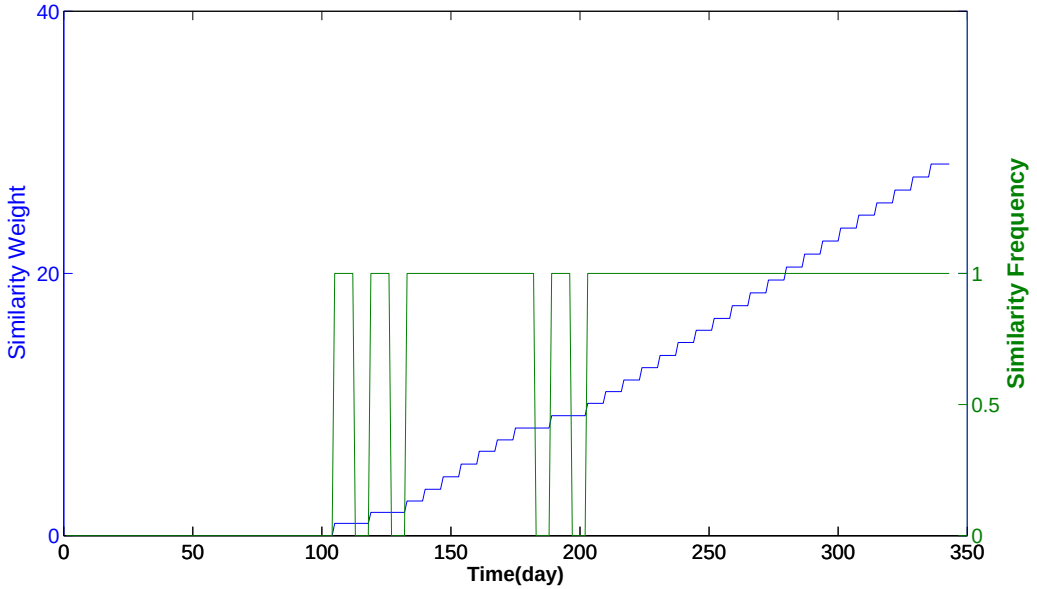



**Figure 7.5:** Two user profiles are not within the same cluster over 49 weeks.

As a counter-example, the similarity frequency $f_{ij}$ of the second user profiles are 0. It denotes that the two user profiles are never within the same cluster during the whole time. Figure 7.5 shows these two user profiles differ widely over 49 weeks. Thus, the similarity frequency $f_{ij}$ can be a good measure to distinguish the pair of user profiles, which are totally different most of the time.




(a) Time-series User Profiles




(b) Similarity Weight and Similarity Frequency

**Figure 7.6:** Two user profiles are within the same cluster over 30 weeks.

As shown in Figure 7.6(a), the front parts of the third pair are very different with each other, but the latter half of the profiles are starting to become similar. The Figure 7.6(b) shows their according similarity weight and similarity frequency with the time changes. The blue line indicates the similarity weight and the green rectangle wave shape indicates whether these two profiles are assigned into the same cluster in one specific week. In Figure 7.6(b), when the similarity weight starts increasing in the latter part, the two raw user profiles are also keeping close. It can be seen that both similarity weight and similarity frequency can be considered as good indicators to show the changing trend of two user profiles.

In this chapter, a segmental clustering algorithm is presented to improve the efficiency on a large scale of real-time smart meter data. It can avoid inaccurate similarity measures that are due to occasional large fluctuations. Another benefit of this algorithm is avoiding the repetitive computing problem exhibited in the regular clustering algorithm when new data appear. We validate this method with three specific examples in which the algorithm is used to trace and identify similar users.

# CHAPTER 8. SUMMARY AND CONCLUSIONS

In this dissertation, the role of in-home behaviors in energy usage analysis is investigated. In particular, we analyze patterns of energy usage by monitoring activities as well as collecting energy usage data from several smart environments. We further identify the role of behaviors for energy usage by using machine learning methods to map activities performed in the environment onto corresponding energy usage. Next, we analyze the energy patterns by identifying frequent sequences of energy usage ranges and identifying outliers in the data. The result of these methods can be used as the basis of an intervention that allows individuals to perform their daily activities in a more energy-efficient manner. All of our algorithms are evaluated based on data collected in the CASAS smart environment testbeds.

To extend our research from individual homes to communities, city-wide smart meter data are studied. To monitor power profiles of each user, a web-based city-wide visualization tool is developed. The tool is beneficial for an analyst to monitor overall energy consumption in large geographic area and to identify potential abnormal usage. We hypothesize that building energy consumption is highly associated with building feature data including building value, size, age, and materials. Thus, machine learning algorithms are leveraged to explore this potential relationship. To

cluster similar user profiles in an efficient manner, longitudinal smart meter data is separated into small pieces over small fixed periods, and then power profiles on each period are clustered respectively. Two similarity factors are designed to estimate the similarity between users in a long-term period. A goal of this study is to validate our hypothesis that energy usage can be analyzed and predicted based on the sensor data that is generated by the residents in a smart home environment. We provide evidence support this hypothesis through empirical validation. Specifically, all of our algorithms and techniques are evaluated using real data collected in actual smart meters in the Pullman area.

The current state-of-the-art approaches for informing residents about their power consumption are limited. They either provide information at too low of a resolution through monthly totals, or at too high of a resolution through instantaneous measurements. The methods and algorithms we introduce put the numbers and costs in context by associating energy with recognizable behaviors. It is this association that can be used to inform people about their impact on the world in a way that they can use to change their own day-to-day activities. The results of this work can be used to provide feedback about a resident's energy consumption as it relates to various activities. Through linking residential behaviors and energy usage, large energy-consuming appliances can be identified. To save energy cost, residents are encouraged to replace these appliances with energy-efficient appliances. By comparing energy-efficient be-

haviors, residents can be notified about which behaviors consume a large amount of energy. For example, residents may forget to turn off or down the heater or air conditioner when they leave the house. Utilizing information we learn from resident behavior and energy consumption, a smart home can remind them to do this on their way out of the house or take the action for them when it detects they have left. In this way, the technologies we design for predicting energy use can form the basis for automating activities in a manner that consumes fewer resources, including power usage. By detecting trends and anomalies, we can find some extreme energy usage values, which may indicate blackout situations, devices that were mistakenly left on, or events that may lead to potential security problems in the smart environment.

Based on our experimental analysis we found that many of these techniques are useful in highlighting data collection issues and behavioral patterns that can affect energy consumption. This link between smart home sensors, machine learning algorithms, and whole-home power usage provide these insights that would otherwise likely not be caught. Additionally, our algorithms are able to show residents how much power their daily behaviors are using. This lays the groundwork for evaluating how effective this information is at influencing activity behavior over time in an attempt to reduce power consumption.

In the area of building sustainability, there are still many challenges that need to be addressed in the future. In the area of smart home research, more sensitive

power meters need to be installed in order to capture more accurate changes in energy consumption. To save cost, a minimally instrumented sensor environment is required to do the same evaluations of our algorithms as are described here. In order to learn more about how building features have an impact on energy usage, the individual features need to be more deeply analyzed. In particular, the features that we identified to be correlated with energy consumption such as building value and land size, can be examined in greater detail to determine the reason for this correlation. Furthermore, the energy-efficiency of building materials needs to be examined more closely by comparison energy usage for similar homes and types of people utilizing the differing materials. By utilizing smart meter data, consumers can be made aware of consumption patterns for similar homes, which in turn can provide insights and motivation to initiate sustainable behaviors. The behavioral interventions can be designed for residents to suggest different ways to perform activities, different timing of activities and automate control of devices to promote energy-efficient behaviors. Furthermore, surveys can be designed and conducted to evaluate the usability of our web-based interface.

Our existing work clusters customer-based similarity measures on aggregated energy usage. In our ongoing work, the hierarchical clusters can be generated to obtain a range of cluster sizes. Given usage information at various population sizes (from individual, to cluster, to community) and temporal sizes (from hour, to day,

to month), we can then design techniques to identify anomalies. We can identify a methodology that can be generalizable to multiple situations. Such situations include identifying time periods that are anomalous for an individual or a group, or identifying individuals/groups that are anomalous over a span of time.

In addition, data mining techniques can be designed to explore whether there is a clear relationship between user demographics (i.e. age, education, occupation and building structure) with raw user profiles. As a result, user power profiles can be taken into account to predict resource consumption. Moreover, in our research, it has been validated that a strong association exists between home energy usage and residential activities. Based on this finding, one promising and challenging area is to explore machine learning methods to recover residential activities by analyzing their daily power profile.

Research on energy-efficient buildings is a challenging area that requires combined efforts in multiple disciplines like electrical engineering, environmental science, computer science, and even psychology. In this dissertation, we concentrate primarily on exploring the relationship between energy consumption and residential behaviors, and designing machine-learning methods to identify patterns and anomalies in energy profiles from both the perspective of individual smart homes and city-wide meter data. We anticipate continued research in these areas for improving building energy efficiency, understanding behavior impact on energy efficiency, and maintaining a

sustainable community.

# A Correlation Coefficient between energy usage and building features.

**Table 8.1:** Correlation coefficient between energy usage and building features (1).

| **Building Features** | **Correlation Coefficient** |
|---|---|
| Building Features | Correlation Coefficient |
| Numbrer of Material Units | 0.380134654 |
| Land Value | 0.368685222 |
| Total Area | 0.36708317 |
| Percent of Ceramic Tile | 0.348721164 |
| Total Value | 0.302305429 |
| Improvement Value | 0.275355464 |
| Pervent of Preformed Metal | 0.235195868 |
| Percent of Masonry Stucco on Block | 0.171899899 |
| Floor Insulation Area | 0.148862471 |
| Level of Building Quality (level 1  level 6) | 0.115160583 |
| Percent of Slab Porch with Roof | 0.065624797 |
| Percent of Light-Weight Concrete | 0.060035884 |

**Table 8.2:** Correlation coefficient between energy usage and building features (2).

| **Building Features** | **Correlation Coefficient** |
|---|---|
| Percent of Frame Siding Metal | 0.057720313 |
| Number of Plumbing Fixtures | 0.05550445 |
| Percent of Veneer Brick | 0.040181568 |
| Type of Land | 0.038059315 |
| Type of unit | 0.034817833 |
| Percent of Electric Radiant Heat | 0.034790652 |
| Percent of Electric Baseboard | 0.027078764 |
| Building Condition (level -1 - level 6) | 0.025223563 |
| Percent of Masonry Face Brick | 0.021459046 |
| Number of Wood Deck | 0.017485618 |
| Percent of Air Condition | 0.015447879 |
| Percent of Frame Plywood | 0.011517937 |
| Percent of Wood Deck with Roof | 0.010062876 |
| Percent of Veneer Stone | 0.008684272 |
| Percent of Masonry Common Brick | 0.00850489 |
| Partition Finish Area | 0.006416652 |
| Percent of Gravity Furnace | 0.005906483 |

**Table 8.3:** Correlation coefficient between energy usage and building features (3).

| **Building Features** | **Correlation Coefficient** |
|---|---|
| Percent of Built-up Rock | 0.003786742 |
| Percent of Rustic Log | 0.003469938 |
| Automatic Floor Cover Allowance | 0.002864359 |
| Automatic Appliance Allowance | 0.002864359 |
| Number of Plumbing Rough-ins | 0.002864359 |
| Percent of Raised Subfloor | 0.002864359 |
| Number of Open Slab Porch | 0.000511598 |
| Percent of Concrete Tile | 0.000264109 |
| Percent of Frame Hardboard | 0.00012647 |
| Outside Entrance Below Grade | 0.00012647 |
| Percent of Wood Stairway | 0.00012647 |
| Percent of Carpet and Pad | -0.000512815 |
| Percent of Wood Shake | -0.000571508 |
| Finished Attached Garage Area | -0.003032505 |
| Number of Lawn Sprinklers | -0.003349928 |
| Percent of Clay Tile | -0.003804212 |
| Percent of Composition Roll | -0.003949664 |

**Table 8.4:** Correlation coefficient between energy usage and building features (4).

| **Building Features** | **Correlation Coefficient** |
|---|---|
| Percent of Frame Stucco | -0.005220613 |
| Percent of Resilient Floor Cover | -0.005656898 |
| Percent of Metal Formed Seams | -0.005741481 |
| Percent of Heat Pump | -0.005822641 |
| Percent of Frame Metal or Vinyl Siding | -0.006641604 |
| Percent of Masonry Concrete Block | -0.006751283 |
| Percent of Masonry Stone on Block | -0.007272703 |
| Percent of Linoleum | -0.008326462 |
| Percent of Oil Fired Boiler | -0.009847058 |
| Percent of Metal Copper | -0.011309793 |
| Percent of Frame Siding Vinyl | -0.012113431 |
| Percent of Floor Radiant Hot Water | -0.012723353 |
| Percent of Baseboard Hot Water | -0.014604532 |
| Percent of Frame Hardboard Sheets | -0.014998568 |
| Percent of Wood Shingle | -0.017064649 |
| Percent of Carport Shed Roof | -0.022184514 |
| Number of Single 1-Story Fireplace | -0.023224236 |

**Table 8.5:** Correlation coefficient between energy usage and building features (5).

| **Building Features** | **Correlation Coefficient** |
|---|---|
| Number of Bedrooms | -0.025115363 |
| Percent of Frame Wood Shingle | -0.025938463 |
| Building Age | -0.027767853 |
| Attached Garage Area | -0.033555904 |
| Percent of Vinyl Sheet | -0.035374327 |
| Number of Bathroom | -0.038968241 |
| Percent of Frame Siding Wood | -0.039987141 |
| Percent of Composition Shingle | -0.042739825 |
| Percent of Forced Air Furnace | -0.044304768 |
| Percent of Hardwood | -0.053810418 |

# Bibliography

Annual energy review 2011. Technical report, U.S. Energy Information Administration, 2012a.

2011 buildings energy data book. Technical report, U.S. Energy Information Administration, 2012b.

Avista company. www.avistacorp.com, 2013a.

Avista smart meter. http://www.avistautilities.com/inside/resources/smartgrid/pages/default.aspx, 2013b.

Aware home research initiative. http://awarehome.imtc.gatech.edu/, 2013.

Duke smart home. http://smarthome.duke.edu/, 2013.

Google Map API. https://developers.google.com/maps/, 2013.

MIT house_n. http://architecture.mit.edu/house_n/, 2013.

Philips HomeLab. http://www.noldus.com/default/philips-homelab, 2013.

Terrascan taxsifter. http://terrascan.whitmancounty.net/Taxsifter, 2013.

Weather underground. http://www.wunderground.com, 2013.

Gediminas Adomavicius and Alexander Tuzhilin. Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. *Knowledge and Data Engineering, IEEE Transactions on*, 17(6):734–749, 2005.

Hunt Allcott. Social norms and energy conservation. *Journal of Public Economics*, 95(9):1082–1095, 2011.

Anil Aswani, Neal Master, Jay Taneja, Andrew Krioukov, David Culler, and Claire Tomlin. Energy-efficient building hvac control using hybrid system lbmpc. In *the IFAC Conference on Nonlinear Model Predictive Control*, 2012a.

Anil Aswani, Neal Master, Jay Taneja, Virginia Smith, Andrew Krioukov, David Culler, and Claire Tomlin. Identifying models of hvac systems using semiparametric regression. In *American Control Conference (ACC)*, pages 3675–3680, 2012b.

Ling Bao and Stephen Intille. Activity recognition from user-annotated acceleration data. *Pervasive Computing*, pages 1–17, 2004.

Gerald Bauer, Karl Stockinger, and Paul Lukowicz. Recognizing the use-mode of kitchen appliances from their current consumption. In *Proceedings of the European conference on Smart sensing and context*, pages 163–176, 2009.

Gowtham Bellala, M. Manish, Martin Arlitt, Geoff Lyon, and C. Bash. Towards an

understanding of campus-scale power consumption. In *ACM Workshop on Embedded Sensing Systems for Energy-Efficiency in Buildings (BuildSys)*, 2011.

Mario Berges, Ethan Goldman, Lucio Soibelman, H. Scott Matthews, and Kyle Anderson. User-centered Non-Intrusive electricity load monitoring for residential buildings. *Journal of Computing in Civil Engineering*, 25(1), 2011.

Donald Berndt and James Clifford. Using dynamic time warping to find patterns in time series. In *Knowledge Discovery and Data Mining*, volume 10, pages 359–370, 1994.

Bernhard E. Boser, Isabelle M. Guyon, and Vladimir N. Vapnik. A training algorithm for optimal margin classifiers. In *Proceedings of the annual workshop on Computational learning theory*, pages 144–152, 1992.

Ronald Bracewell. *The fourier transform & its applications 3rd Ed.* McGraw-Hill Science/Engineering/Math, 2000.

Oliver Brdiczka, Patrick Reignier, and James Crowley. Detecting individual activities from video in a smart home. In *Knowledge-Based Intelligent Information and Engineering Systems*, pages 363–370, 2007.

Nitesh V. Chawla, Kevin W. Bowyer, Lawrence O. Hall, and W. Philip Kegelmeyer.

SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16:321–357, 2002.

Chao Chen and Diane Cook. Energy outlier detection in smart environments. In *Workshops at the Twenty-Fifth AAAI Conference on Artificial Intelligence*, 2011a.

Chao Chen and Diane J. Cook. Novelty detection in human behavior through analysis of energy utilization. In *Human Behavior Recognition Technologies, IGI Global*. 2011b.

Chao Chen and Diane J. Cook. Behavior-based home energy prediction. In *2012 8th International Conference on Intelligent Environments (IE)*, pages 57–63, 2012.

Chao Chen and Prafulla Dawadi. CASASviz: Web-based visualization of behavior patterns in smart environments. In *IEEE International Conference on Pervasive Computing and Communications*, pages 650–652, 2011.

Chao Chen, Barnan Das, and Diane J. Cook. Energy prediction based on residents activity. In *Proceedings of the International workshop on Knowledge Discovery from Sensor Data*, 2010a.

Datong Chen, Jie Yang, and Howard Wactlar. A study of detecting social interaction with sensors in a nursing home environment. *Computer Vision in Human-Computer Interaction*, pages 199–210, 2005.

Liming Chen, Chris D Nugent, and Hui Wang. A knowledge-driven approach to activity recognition in smart homes. *IEEE Transactions on Knowledge and Data Engineering*, 24(6):961–974, 2012.

Ming-yu Chen, Lily Mummert, Padmanabhan Pillai, Alex Hauptmann, and Rahul0 Sukthankar. Exploiting multi-level parallelism for low-latency activity recognition in streaming video. In *Proceedings of the first annual ACM SIGMM conference on Multimedia systems*, page 112, 2010b.

Gabe Cohn, Sidhant Gupta, Jon Froehlich, Eric Larson, and Shwetak N. Patel. GasSense: appliance-level, single-point sensing of gas activity in the home. *Pervasive Computing*, pages 265–282, 2010.

Crandall Aaron S. Thomas Brian L. Krishnan Narayanan C. Cook, Diane J. CASAS: A smart home in a box. *IEEE Computer*, to appear.

Diane J. Cook and Sajal K. Das. *Smart environments: technologies, protocols, and applications*. Wiley Series on Parallel and Distributed Computing. Wiley-Interscience, 2004.

Diane J. Cook, Michael Youngblood, Edwin O. Heierman, Karthik Gopalratnam, Sira Rao, Andrey Litvin, and Farhan Khawaja. MavHome: an agent-based smart home.

In *Proceedings of the IEEE International Conference on Pervasive Computing and Communications*, pages 521–524, 2003.

Diane J. Cook, Maureen Schmitter-Edgecombe, Aaron Crandall, Chad Sanders, and Brian Thomas. Collecting and disseminating smart home sensor data in the CASAS project. In *Proceedings of the CHI Workshop on Developing Shared Home Behavior Datasets to Advance HCI and Ubiquitous Computing Research*, 2009.

James W. Cooley and John W. Tukey. An algorithm for the machine calculation of complex fourier series. *Mathematics of Computation*, 19(90):297, 1965.

I. D. Coope. Circle fitting by linear and nonlinear least squares. *Journal of Optimization Theory and Applications*, 76(2):381–388, 1993.

Aaron S. Crandall and Diane J. Cook. Tracking systems for multiple smart home residents. *Human Behavior Recognition Technologies, IGI Global*, 2011.

Dave Crane and Phil McCarthy. *Comet and Reverse Ajax: The Next Generation Ajax 2.0.* Apress, 2008.

Teri Crosby. How to detect and handle outliers. *Technometrics*, 36(3):315–316, 1994.

Sarah Darby. The effectiveness of feedback on energy consumption. *A Review for DEFRA of the Literature on Metering, Billing and direct Displays*, 486, 2006.

Sarah Darby. Smart metering: what potential for householder engagement? *Building Research & Information*, 38(5):442–457, 2010. doi: 10.1080/09613218.2010.492660.

Barnan Das, Diane J. Cook, Maureen Schmitter-Edgecombe, and Adriana M. Seelye. PUCK: an automated prompting system for smart environments. *Personal and Ubiquitous Computing Theme Issue on Sensor-driven Computing and Applications for Ambient Intelligence*, 2012a.

Barnan Das, Adriana M. Seelye, Brian L. Thomas, Diane J. Cook, Larry B. Holder, and Maureen Schmitter-Edgecombe. Using smart phones for context-aware prompting in smart environments. In *2012 IEEE Consumer Communications and Networking Conference*, pages 399–403, 2012b.

W. Edwards Deming. On probability as a basis for action. *American Statistician*, 29 (4):146–152, 1975.

Colin Dixon, Ratul Mahajan, Sharad Agarwal, AJ Brush, Bongshin Lee, Stefan Saroiu, and Victor Bahl. An operating system for the home. In *Proceedings of the 9th USENIX Symposium on Networked Systems Design and Implementation*, 2012.

Jon Froehlich, Eric Larson, Tim Campbell, Conor Haggerty, James Fogarty, and Shwetak N. Patel. HydroSense: infrastructure-mediated single-point sensing of

whole-home water activity. In *Proceedings of the 11th international conference on Ubiquitous computing*, pages 235–244, 2009.

Lei Gao, Alan K Bourke, and John Nelson. Activity recognition using dynamic multiple sensor fusion in body sensor networks. In *Proceedings of the annual international conference of the IEEE Engineering in Medicine and Biology Society*, pages 1077–1080. IEEE, 2012.

Hassan Ghasemzadeh and Roozbeh Jafari. Physical movement monitoring using body sensor networks: A phonological approach to construct spatial decision trees. *IEEE Transactions on Industrial Informatics*, 7(1):66–77, 2011.

B. Griffith and D. Crawley. *Methodology for Analyzing the Technical Potential for Energy Performance in the US Commercial Buildings Sector with Detailed Energy Modeling: Preprint.* National Renewable Energy Laboratory, 2006.

Tao Gu, Zhanqing Wu, Xianping Tao, Hung Keng Pung, and Jian Lu. epsicar: An emerging patterns based approach to sequential, interleaved and concurrent activity recognition. In *Proceedings of the IEEE International Conference on Pervasive Computing and Communications*, pages 1–9, 2009.

Eric Guenterberg, Hassan Ghasemzadeh, and Roozbeh Jafari. Automatic segmenta-

tion and recognition in body sensor networks using a hidden markov model. *ACM Transactions on Embedded Computing Systems (TECS)*, 11(S2):46, 2012.

Steve R Gunn. Support vector machines for classification and regression. *ISIS technical report*, 14, 1998.

Sidhant Gupta, Matthew S. Reynolds, and Shwetak N. Patel. Electrisense: single-point sensing using emi for electrical event detection and classification in the home. In *Proceedings of the 12th ACM international conference on Ubiquitous computing*, pages 139–148, 2010.

Dan Gusfield. *Algorithms on strings, trees, and sequences: computer science and computational biology*. Cambridge University Press, 1997.

Robert Gwadera, Mikhail J. Atallah, and Wojciech Szpankowski. Markov models for identification of significant episodes. In *Proceedings of the 5th SIAM international conference on data mining*, pages 404–414, 2005.

Colin Harris and Vinny Cahill. Exploiting user behaviour for context-aware power management. In *In Proceedings of the IEEE International Conference on Wireless And Mobile Computing, Networking And Communications*, pages 122–130, 2005.

Colin Harris and Vinny Cahill. An empirical study of the potential for context-aware power management. *UbiComp 2007: Ubiquitous Computing*, pages 235–252, 2007.

George W. Hart. Nonintrusive appliance load monitoring. *Proceedings of the IEEE*, 80(12):1870–1891, 1992.

John A. Hartigan. *Clustering algorithms*. John Wiley & Sons, Inc., 1975.

Sumi Helal, Willian Mann, Hicham El-Zabadani, Jeffrey King, Youssef Kaddoura, and Erwin Jansen. The gator tech smart house: A programmable pervasive space. *Computer*, page 50–60, 2005.

Toshiaki Ichinose, Kazuhiro Shimodozono, and Keisuke Hanaki. Impact of anthropogenic heat on urban climate in tokyo. *Atmospheric Environment*, 33(24-25): 3897–3909, 1999.

Koichiro Ito. The effect of financial incentives on energy conservation: A regression discontinuity design in the california 20/20 program. *University of California, Berkeley*, 2010.

Vikramaditya Jakkula and Diane J. Cook. Outlier detection in smart environment structured power datasets. In *Proceedings of the 6th International Conference on Intelligent Environments (IE),*, pages 29 –33, 2010.

Xiaofan Jiang, Stephen Dawson-Haggerty, Prabal Dutta, and David Culler. Design and implementation of a high-fidelity ac metering network. In *International Conference on Information Processing in Sensor Networks*, pages 253–264, 2009.

Paul L Joskow and Catherine D Wolfram. Dynamic pricing of electricity. *The American Economic Review*, 102(3):381–385, 2012.

Karen Kafadar. Statistical process control: The deming paradigm and beyond. *Technometrics*, 45(1):103–104, 2003.

Takekazu Kato, Hyun Cho, Dongwook Lee, Tetsuo Toyomura, and Tatsuya Yamazaki. Appliance recognition from electric current signals for Information-Energy integrated network in home environments. *Ambient Assistive Health and Wellness Management in the Heart of the City*, page 150–157, 2009.

Eamonn Keogh, Jessica Lin, Sang-Hee Lee, and Helga Herle. Finding the most unusual time series subsequence: algorithms and applications. *Knowledge and Information Systems*, 11(1):1–27, 2007.

Eunju Kim, Sumi Helal, and Diane J. Cook. Human activity recognition and pattern discovery. *IEEE Pervasive Computing*, 9(1):48–53, 2010.

J. Zico Kolter and Joseph Ferreia Jr. A large-scale study on predicting and contextualizing building energy usage. In *the Conference on Artificial Intelligence (AAAI), Special Track on Computational Sustainability and AI*, 2011.

Ivan Korolija, Ljiljana Marjanovic-Halburd, Yi Zhang, and Victor I Hanby. Uk office buildings archetypal model as methodological approach in development of regres-

sion models for predicting building energy consumption from heating and cooling demands. *Energy and Buildings*, 60:152–162, 2013.

James Kusznir. CLM as a smart home middleware. Master's thesis, Washington State University, 2010.

Walter S Lasecki, Young Chol Song, Henry Kautz, and Jeffrey P Bigham. Real-time crowd labeling for deployable activity recognition. In *Proceedings of the 2013 conference on Computer supported cooperative work*, pages 1203–1212. ACM, 2013.

Yinhong Li, H-D Chiang, B-K Choi, Y-T Chen, D-H Huang, and Mark G Lauby. Representative static load models for transient stability analysis: development and examination. *Generation, Transmission & Distribution, IET*, 1(3):422–431, 2007.

Jessica Lin, Eamonn Keogh, Ada Fu, and Helga Van Herle. Approximations to magic: finding unusual medical time series. In *Proceedings of the 18th IEEE Symposium on Computer-Based Medical Systems*, pages 329–334, 2005.

Greg Linden, Brent Smith, and Jeremy York. Amazon.com recommendations: Item-to-item collaborative filtering. *Internet Computing, IEEE*, 7(1):76–80, 2003.

Huan Liu, Farhad Hussain, Chew Lim Tan, and Manoranjan Dash. Discretization: An enabling technique. *Data Mining and Knowledge Discovery*, 6(4):393–423, 2002.

Kantilal Varichand Mardia, John T. Kent, and John M. Bibby. *Multivariate Analysis*. Academic Press, 1980.

Uwe Maurer, Anthony Rowe, Asim Smailagic, and Daniel Siewiorek. Location and activity recognition using eWatch: a wearable sensor platform. *Ambient Intelligence in Everyday Life*, pages 86–102, 2006a.

Uwe Maurer, Asim Smailagic, Daniel P Siewiorek, and Michael Deisher. Activity recognition and monitoring using multiple sensors on different body positions. In *International Workshop onWearable and Implantable Body Sensor Networks*, page 4, 2006b.

Edward M. McCreight. A space-economical suffix tree construction algorithm. *Journal of the ACM*, 23(2):262–272, 1976.

J. Michael McQuade. A system approach to high performance buildings. *United Technologies Corporation, Tech. Rep*, 2009.

Jan Meyer, Paul Lukowicz, and Gerhard Troster. Textile pressure sensor for muscle activity and motion detection. In *Proceedings of 10th IEEE International Symposium on Wearable Computers*, pages 69–72. IEEE, 2006.

G. Mihalakakou, M. Santamouris, and A. Tsangrassoulis. On the energy consumption in residential buildings. *Energy and buildings*, 34(7):727–736, 2002.

Paresh Kumar Narayan and Russell Smyth. The residential demand for electricity in australia: an application of the bounds testing approach to cointegration. *Energy Policy*, 33(4):467–474, 2005.

Mark EJ Newman. Power laws, pareto distributions and zipf's law. *Contemporary physics*, 46:323–351, 2005.

Guy R. Newsham and Benjamin J. Birt. Building-level occupancy data to improve ARIMA-based electricity use forecasts. In *Proceedings of the 2nd ACM Workshop on Embedded Sensing Systems for Energy-Efficiency in Building*, pages 13–18, 2010.

Xiufeng Pang, Prajesh Bhattacharya, Zheng ONeill, Philip Haves, Michael Wetter, and Trevor Bailey. Real-time building energy simulation using EnergyPlus and the building controls virtual test bed. *Proceedings of Building Simulation 11*, 2011.

Alexandros Pantelopoulos and Nikolaos G. Bourbakis. A survey on wearable sensor-based systems for health monitoring and prognosis. *IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews*, 40(1):1–12, 2010.

Judea Pearl. *Probabilistic reasoning in intelligent systems: networks of plausible inference*. Morgan Kaufmann, 1988.

Hanchuan Peng, Fuhui Long, and Chris Ding. Feature selection based on mutual in-

formation: criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Transactions on pattern analysis and machine intelligence*, 27(8), 2005.

Clifton Phua, Damminda Alahakoon, and Vincent Lee. Minority report in fraud detection: classification of skewed data. *ACM SIGKDD Explorations Newsletter*, 6 (1):50–59, 2004.

John Platt. Sequential minimal optimization: A fast algorithm for training support vector machines. Technical report, Microsoft Research, 1998.

Thomas A. Powell. *Ajax: The Complete Reference*. McGraw Hill Professional, 2008.

Juhi Ranjan, Yu Yao, Erin Griffiths, and Kamin Whitehouse. Using mid-range rfid for location based activity recognition. In *Proceedings of the 2012 ACM Conference on Ubiquitous Computing*, pages 647–648. ACM, 2012.

Parisa Rashidi, Diane J. Cook, Lawrence B. Holder, and Maureen Schmitter-Edgecombe. IEEE transactions on discovering activities to recognize and track in a smart environment. *Knowledge and Data Engineering*, 23(4):527–539, 2011.

Christian Reinisch, Mario J Kofler, Félix Iglesias, and Wolfgang Kastner. Thinkhome energy efficiency in future smart homes. *EURASIP Journal on Embedded Systems*, 2011:1, 2011.

John Rice. Mathematical statistics and data analysis. *Duxbury Press*, 2006.

Bernard Rosner. Percentage points for a generalized ESD many-outlier procedure. *Technometrics*, 25(2):165–172, 1983.

Peter J. Rousseeuw. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, 20:53–65, 1987.

Abhishek Roy, Soumya K. Das Bhaumik, Amiya Bhattacharya, Kalyan Basu, Diane J. Cook, and Sajal K. Das. Location aware resource management in smart homes. In *Proceedings of the 1st IEEE International Conference on Pervasive Computing and Communications*, pages 481–488, 2003.

Michael S. Ryoo and Jake K. Aggarwal. Spatio-temporal relationship match: Video structure comparison for recognition of complex human activities. In *Proceedings of the 12th IEEE International Conference on Computer Vision*, pages 1593–1600, 2009.

David J Sailor and Lu Lu. A top-down methodology for developing diurnal and seasonal anthropogenic heating profiles for urban areas. *Atmospheric Environment*, 38(17):2737–2748, 2004.

John E. Seem. Using intelligent data analysis to detect abnormal energy consumption in buildings. *Energy and Buildings*, 39(1):52–58, January 2007.

John Seryak and Kelly Kissock. Occupancy and behavioral affects on residential energy use. In *Proceedings of the Solar Conference*, pages 717–722, 2003.

Geetika Singla, Diane J. Cook, and Maureen Schmitter-Edgecombe. Tracking activities in complex settings using smart environment technologies. *International journal of biosciences, psychiatry, and technology*, 1(1):25, 2009.

Xiaoyuan Su and Taghi M. Khoshgoftaar. A survey of collaborative filtering techniques. *Advances in Artificial Intelligence*, page 4, 2009.

S. Szewcyzk, K. Dwan, B. Minor, B. Swedlove, and D. Cook. Annotating smart environment sensor data for activity learning. *Technology and Health Care*, 17(3): 161–169, 2009.

Emmanuel Tapia, Stephen Intille, and Kent Larson. Activity recognition in the home using simple and ubiquitous sensors. *Pervasive Computing*, pages 158–175, 2004.

Andreas Tscher, Michael Jahrer, and Robert Legenstein. Improved neighborhood-based algorithms for large-scale recommender systems. In *Proceedings of the 2nd KDD Workshop on Large-Scale Recommender Systems and the Netflix Prize Competition*, page 4, 2008.

John W Tukey. Exploratory data analysis. *Reading, MA*, 231, 1977.

Tim van Kasteren and Ben Krose. Bayesian activity recognition in residence for elders. In *Proceedings of the 3rd IET International Conference on Intelligent Environments*, pages 209–212, 2007.

Grayson K Vincent and Victoria Averil Velkoff. The next four decades the older population in the united states: 2010 to 2050. *US Census Bureau*, (1138), 2010.

Sanford Weisberg. *Applied linear regression*. Wiley, 2005.

Charlie Wilson and Hadi Dowlatabadi. Models of decision making and residential energy use. *Annual Reviews Environment Resources*, 32:169–203, 2007.

Ian H. Witten and Eibe Frank. *Data mining: practical machine learning tools and techniques*. Morgan Kaufmann, 2005.

Jaeyoung Yang, Joonwhan Lee, and Joongmin Choi. Activity recognition based on RFID object usage for smart mobile devices. *Journal of Computer Science and Technology*, 26(2):239–246, 2011.

Jin Yang, Hugues Rivard, and Radu Zmeureanu. On-line building energy prediction using adaptive artificial neural networks. *Energy and Buildings*, 37(12):1250–1259, December 2005.

Runming Yao and Koen Steemers. A method of formulating energy load profile for domestic buildings in the UK. *Energy and Buildings*, 37(6):663–671, 2005.

Kejia Zhang, Shengfei Shi, Hong Gao, and Jianzhong Li. Unsupervised outlier detection in sensor networks using aggregation tree. *Advanced Data Mining and Applications*, pages 158–169, 2007.

M Zia Uddin, JJ Lee, and T-S Kim. Independent shape component-based human activity recognition via hidden markov model. *Applied Intelligence*, 33(2):193–206, 2010.

Steven F. Zornetzer. *An introduction to neural and electronic networks.* Morgan Kaufmann, 1995.