# Novelty Detection in Human Behavior through Analysis of Energy Utilization

**Chao Chen and Diane J. Cook**
*School of Electrical Engineering and Computer Science, Washington State University, USA*

## ABSTRACT

The value of smart environments in understanding and monitoring human behavior has become increasingly obvious in the past few years. Using data collected from sensors in these environments, scientists have been able to recognize activities that residents perform and use the information to provide context-aware services and information. However, less attention has been paid to monitoring and analyzing energy usage in smart homes, despite the fact that electricity consumption in homes has grown dramatically. In this chapter we demonstrate how energy consumption relates to human activity through verifying that energy consumption can be predicted based on the activity that is being performed. We then automatically identify novelties in human behavior by recognizing outliers in energy consumption generated by the residents in a smart environment the reasons for these abnormalities. To validate these approaches, we use real energy data collected in our CASAS smart apartment testbed and analyze the results for three different data sets collected in this smart home.

## INTRODUCTION

Smart homes have become a very popular research area. One of the most exciting applications of this work is activity recognition and health monitoring for homes. Smart homes provide a forum for observing how these activities are performed, how they are affected by a variety of conditions, and for better understanding the nature of human behavior. Most of the analyzed sensors are used to identify the location of the environment residents as well as the objects with which they are interacting.
In this book chapter we focus on a different type of information that can be gathered and analyzed in smart environments. In particular, we gather and analyze electricity usage that is generated in a smart environment. By observing energy consumption in a smart environment we can perform novel types of analyses that correlate activity performance with energy consumption. We can also analyze energy consumption by itself to detect anomalies in the data and see if they correlate back to abnormalities in resident behavior.

The long-term vision for this project is to enhance understanding of human resource consumption and to provide resource efficiency in smart homes. We envision this as a three step process: 1) predict the energy that will is used to support specific daily activities, 2) analyze electricity usage to identify trends and anomalies, and 3) automate activity support in a more energy-efficient manner. This chapter addresses the first two steps in the process. We hypothesize that energy consumption is correlated with the type of activities that are performed and can therefore be predicted based on the automatically-recognized activities that occur in a smart environment. We further postulate that anomalies can be automatically detected and that these outliers can provide inside on novelties that occur in the behavior of residents in the space. We validate these hypotheses by implementing algorithms to perform these steps and

evaluating the algorithms using data collected in the CASAS smart apartment testbed. The result of this work can be used to give smart home residents feedback on energy consumption.
In the next section of the chapter we summarize related work in the area of smart homes and activity recognition. In the following section we introduce our CASAS smart environment architecture and describe our data collection modules as well as the smart apartment testbed. We next describe our method of predicting energy consumption based on the activity that is performed in the smart environment and evaluate the algorithm using CASAS datasets. After this, we describe the main statistical methods we utilize to detect outliers in energy usage and validate the approach using two different smart home energy data sets.

## BACKGROUND

Given the recent progress in computing power, networking, and sensor technology, we are steadily moving into the world of ubiquitous computing where technology recedes into the background of our lives. Using sensor technology combined with the power of data mining and machine learning, many researchers are now working on smart environments which can discover and recognize residents' activities and respond to resident needs in a context-aware way.

A core technology component in this research is the ability to automatically recognize and identify activities performed by residents in smart environments. A variety of approaches have been used to achieve this goal. For example, Hu et al. [1] find common trends in Activities of Daily Living (ADLs) to see whether the inhabitants perform multiple concurrent and interleaved activities or single activities. Gao et al. [2] use hidden Markov models to characterize different stages in dining activities. The smart hospital project [3] develops a robust approach for recognizing user's activities and estimating hospital-staff activities by employing a hidden Markov model with contextual information in the smart hospital environment. The Georgia Tech Aware Home [4] identifies people based on pressure sensors embedded into the smart floor in strategic locations. The CASAS smart home project [5] builds probabilistic models of activities and uses them to recognize activities in complex situations where multiple residents perform activities in parallel in the same environment. A new idea of transfer learning [6] is gaining popularity in smart home research due to its ability to use the knowledge gained from one domain to a different but related domain, making the learning problem more generalized for similar environments, activities, or inhabitants.

We note that these projects focus primarily on activity recognition using sensors for motion and object interaction. However, very few projects are expanding their scope to consider the resource utilization of smart home residents. Based on a recent report [7], buildings are responsible for at least 40% of energy use in most countries. Furthermore, household consumption of electricity has been growing dramatically. Thus, the need to develop technologies that improve energy efficiency and monitor the energy usage of inhabitants in a household is emerging as a critical research area. The BeAware project [8] makes use of an iPhone application to give users alerts and to provide information on the energy consumption of the entire house. This mobile application can detect the electricity consumption of different devices and notify the user if the devices use more energy than expected. The PowerLine Positioning (PLP) indoor location system [9] is able to localize to sub-room level precision by using fingerprinting of the amplitude of tones produced by two modules installed in extreme locations of the home. Patel et al [10] records and analyzes electrical noise on the power line caused by the switching of significant electrical loads by a single, plug-in module, which can connect to a personal computer, then uses machine learning techniques to identify unique occurrences of switching events (events which denote a change in the status of an electrical device) by tracking the patterns of electrical noise. The MITes platform [11] monitors changes of various appliances in current electricity flow for the appliance, such as a switch from on to off by installing current sensors for each appliance. Other similar work [12] also proposes several approaches to

recognize the energy usage of electrical devices by the analysis of a power line current. These can detect whether the appliance is used and how it is used.

## CASAS SMART ENVIRONMENT

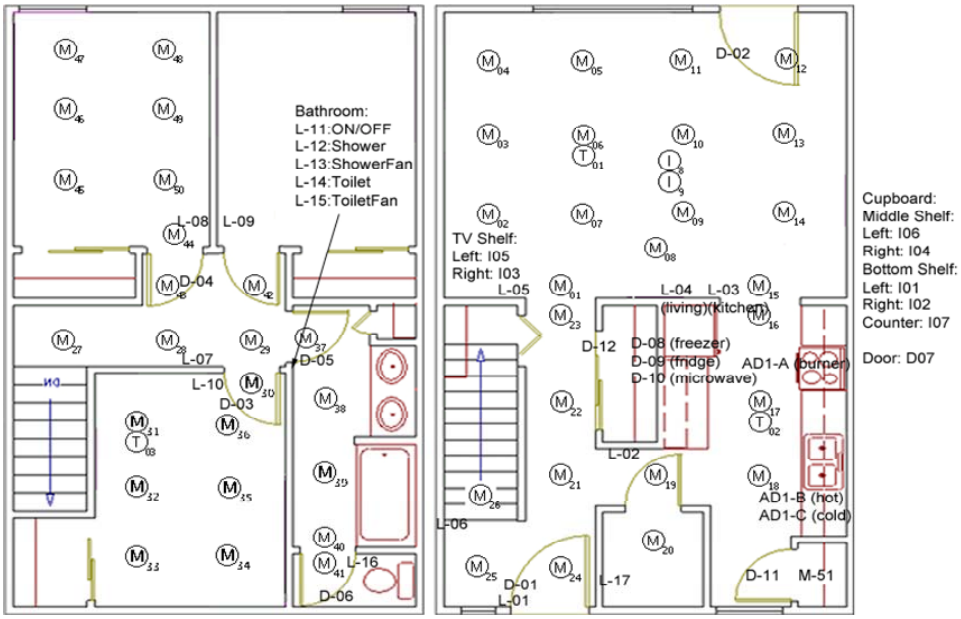


Figure 1. Three-bedroom smart apartment used for our data collection (motion (M), temperature (T), water (W), burner (B), telephone (P),and item (I)).

The smart home environment test bed that we are using to recognize the status of each device is a three bedroom apartment located on the Washington State University campus. As shown in Figure 1, the smart home apartment testbed consists of three bedrooms, one bathroom, a kitchen, and a living/dining room. To track people's mobility, we use motion sensors placed on the ceilings. The circles in the figure stand for the positions of motion sensors. They facilitate tracking the residents who are moving through the space. In addition, the test bed 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 in real time.

The sensor data gathered for our SQL database is expressed by several features, summarized in Table 1. These four fields (Date, Time, Sensor, ID and Message) are generated by the CASAS data collection system automatically.

**Table 1.** Raw Data from Sensors

| Date | Time | Sensor ID | Message |
|---|---|---|---|
| 2009-02-06 | 17:17:36 | M45 | ON |
| 2009-02-06 | 17:17:40 | M45 | OFF |
| 2009-02-06 | 11:13:26 | T004 | 21.5 |
| 2009-02-06 | 11:18:37 | P001 | 747W |
| 2009-02-09 | 21:15:28 | P001 | 1.929kWh |

To provide real training data, we have collected data while two students in good health were living in the smart apartment. Our training data was gathered during several months and more than 100,000 sensor events were generated during this time. Each student had a separate bedroom and shared the downstairs living areas in the smart apartment. All of our experimental data are produced by the day to day lives of these students, which guarantee that the results of this analysis are real and useful.

After collecting data from the CASAS smart apartment, we annotated the sensor events with the corresponding activities that were being performed while the sensor events were generated. Because the annotated data is used to train the machine learning algorithms, the quality of the annotated data is very important for the performance of the learning algorithms. As a large number of sensor data events are generated in a smart home environment, it becomes difficult for researchers and users to convert sequence of sensor events into descriptions of resident activities [20] without the use of visualization tools.

To improve the quality of the annotated data, we built an open source Python-based sensor event visualize, called PyViz, to graphically display the sensor events. Figure 2 shows a screenshot of PyViz as it shows sensor events occurring in the CASAS smart apartment. We also make use of PyViz's Annotation Visualizer to graph recognized resident activities, as shown in Figure 3.

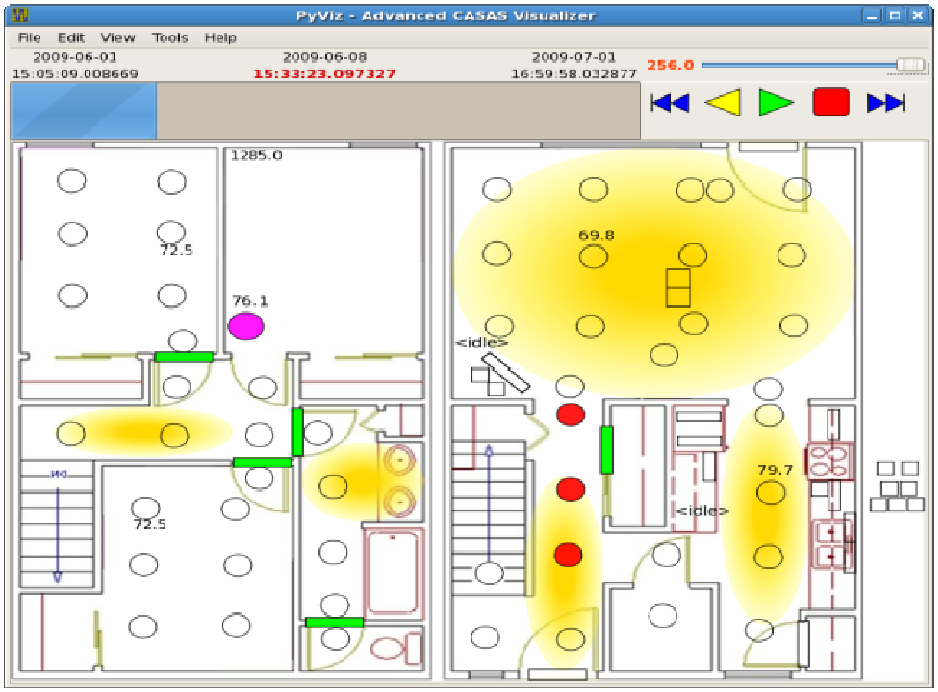


Figure 2. PyViz visualizer.

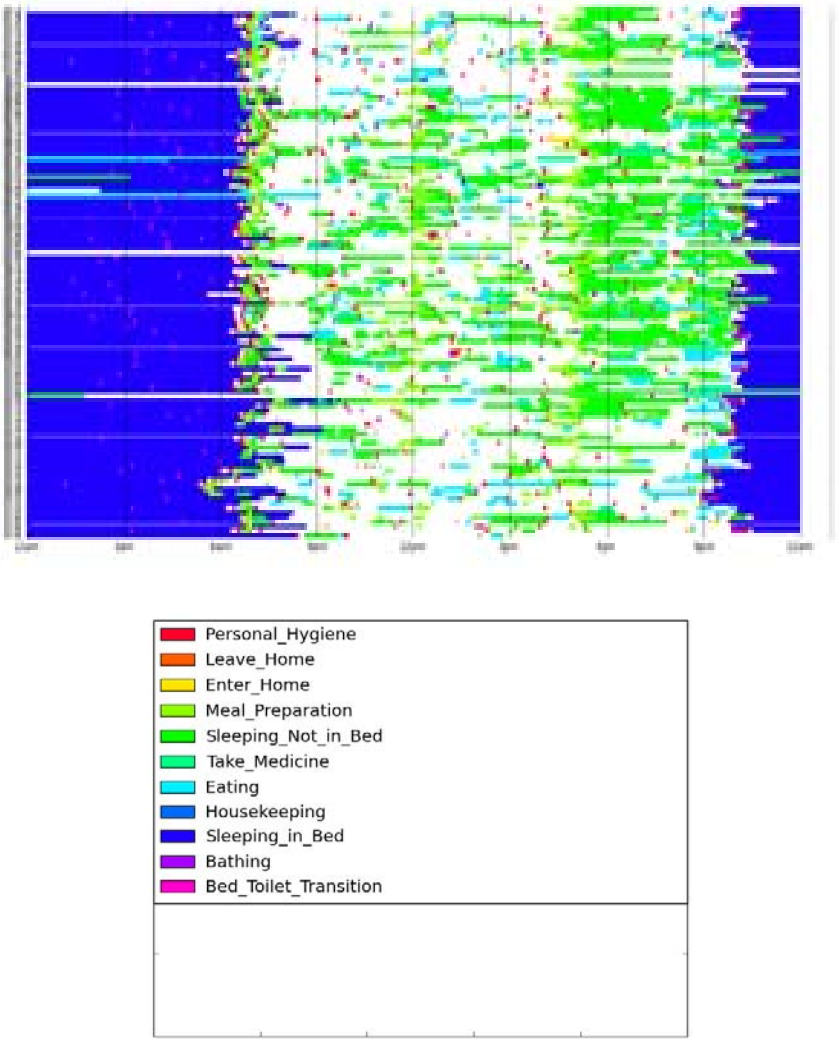


Figure 3. Visualizing activities in a smart home environment.

With the help of PyViz, activity labels are optionally added to each sensor event, providing a label for the current activity. For our experiment, we selected six activities that the two volunteer participants regularly perform in the smart apartment to predict energy use. These activities are as follows:

1. Work at computer
2. Sleep
3. Cook
4. Watch TV
5. Shower
6. Groom

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 2 gives a list of appliances associated with each activity. It should be noted that, there are some appliances which are in "always on" mode, 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 2. Electrical appliances associated with each activity.

| **Activity** | **Appliances Directly Associated** | **Appliances Indirectly Associated** |
|---|---|---|
| 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 |
| Groom | Blow drier | Localized lights |

## ENERGY ANALYSIS

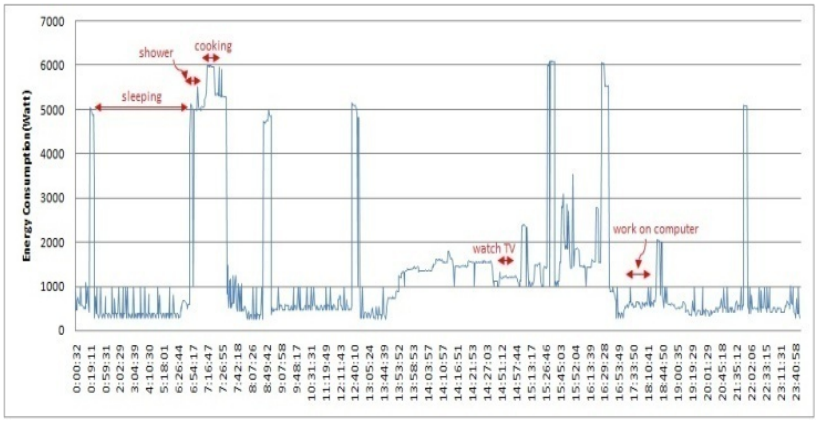


Figure 4. Energy usage for a single day.

Figure 4 shows the energy fluctuation that occurred during a single day on June 2nd, 2009. The activities have been represented by red arrows. The length of the arrows indicates the duration of time (not to scale) for different activities. 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 appliances, even though it was not used directly. The water heater starts heating by itself whenever the temperature of water falls below a certain threshold.

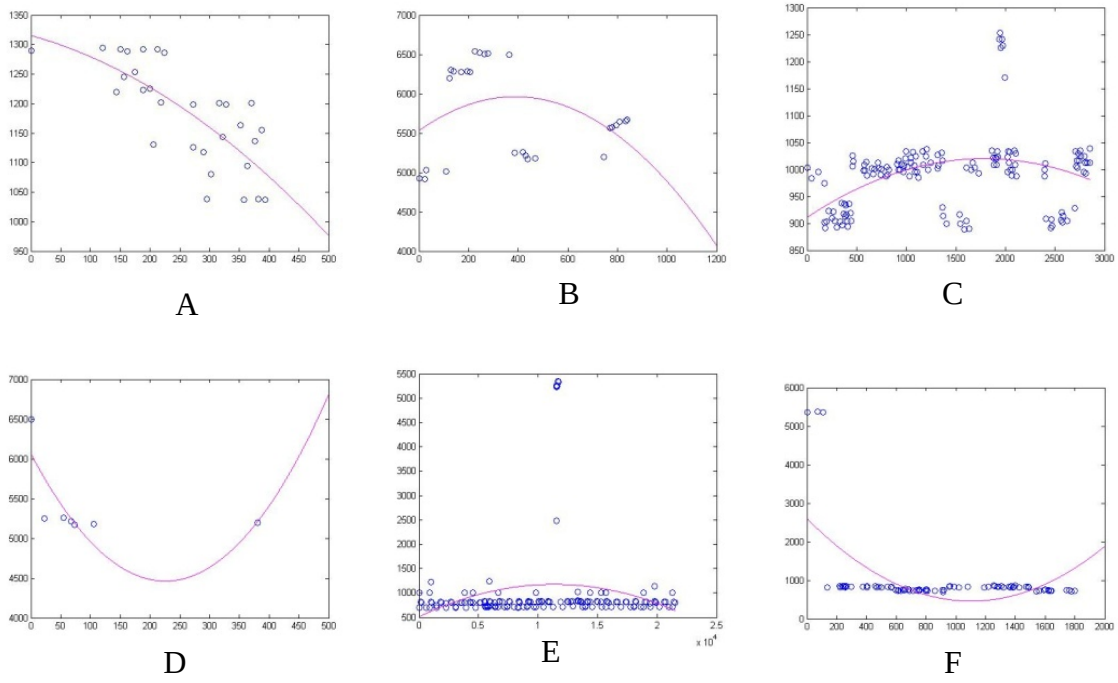


Figure 5. Energy data curve fitting for each activity. There is a separate graph for each activity: A=shower, B=cook, C=work on computer, D=groom, E=sleep, and F=watch TV. The x axis in the graphs represents wattage and the y axis represents time of the activity in seconds.

Figure 5 plots typical energy data for each activity together with the result of applying curve fitting to the data. Curve fitting [14] 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, to approximate the values when no data are available, and to express the relationships between different data points. From the figure, we see that each resident's activity generates different energy patterns. The “cook” activity consumes the highest energy because the participants may open the refrigerator and use the stove or microwave oven, which need a relatively high power. Meantime, when the participants were sleeping, the energy consumption was the lowest because most appliances were idle.

## ENERGY PREDICTION

In the first step of our goal, we use machine learning techniques to predict energy consumption given information about an activity that inhabitants perform in a smart environment. We use the following features to describe an activity performed by an inhabitant in a smart home:

1. Activity label
2. Activity length, measured in seconds
3. Previous activity
4. Next activity
5. Number of different motion sensors fired during activity
6. Total number of motion sensor events
7. Motion sensor On/Off settings for each motion sensor in the space

We use several machine learning algorithms to map these activity features onto a label indicating the amount of energy that is consumed in the smart environment while the activity was performed. In this study, we make use of four popular machine learning methods: a naïve Bayes classifier (NBC), a Bayes net classifier (BNC), a neural network (NN), and a support vector machine (SVM). A naïve Bayes classifier is a probabilistic classifier that assumes the presence of a particular input feature is unrelated to any of the other features given the target label. This classifier applies Bayes theorem to learn a mapping from the input features to the classification label.

$$\text{argmax}_{e_i \in E} P(e_i|F) = \frac{P(F|e_i)P(e_i)}{P(F)} \quad (1)$$

In Equation 1, *E* represents the energy class label and *F* represents the input features described above. The value of $P(e_i)$ is estimated based on the relative frequency with which each target value $e_i$ occurs in the training data. Based on the simplifying assumption that feature values are independent given the target value, the probabilities of observing the particular data point (activity) is the product of the probabilities of the individual features describing the activity, calculated using Equation 2.

$$P(F|e_j) = \prod_i P(f_i|e_j) \quad (2)$$

Bayes belief networks [20] belong to the family of probabilistic graphical models. They 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 by the variable name. Unlike the naïve 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 the 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.

Artificial Neural Networks (ANNs) [21] are abstract computational models based on the organizational structure of the human brain. ANNs provide a general and robust method to learn a target function from input examples. The most common learning method for ANNs, called Backpropagation, which performs a gradient descent within the solution's 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 choose the Multilayer-Perceptron algorithm with Backpropagation to predict electricity usage.

Super Vector Machines (SVMs) were first introduced in 1992 [22]. This is a training algorithm for data classification, which maximizes the margin between the training examples and the class boundary. The SVM learns a hyperplane which separates instances from multiple activity classes with maximum margin.

Each training data instance should contain one class label and several features. The goal of a SVM is to generate a hyperplane which provides a class label for each data point described by a set of feature values.

## Experimental Results

We performed two series of energy prediction experiments. The first experiment uses the sensor data collected during two summer months in the testbed. In the second experiment, we collected data of three winter months in the testbed. The biggest difference between these two groups of data is that some high energy consuming devices like room heaters were only used during the winter, which are not directly controlled by the residents and are therefore difficult to monitor and predict. Using the Weka machine learning toolset [23], we assessed the classification accuracy of our four selected machine learning algorithms and reported the predictive accuracy results based on a 3-fold cross validation.

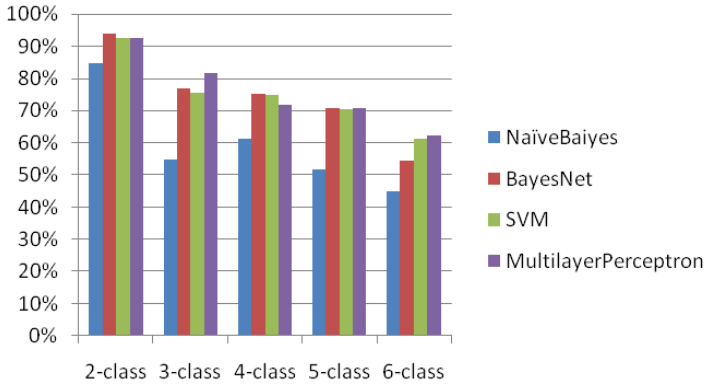


Figure 6. Comparison of the accuracy for summer dataset.

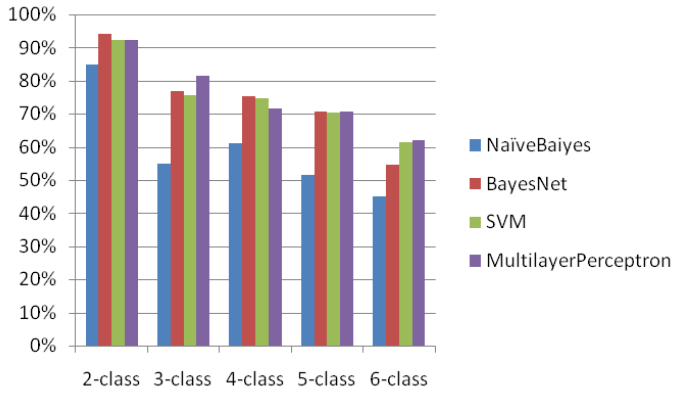


Figure 7. Comparison of the accuracy for winter dataset.

Figures 6 and 7 plot the accuracies of the two different group experiments, respectively. As shown in these two figures, the highest accuracy is around 90% for both datasets to predict the two-class energy usage and the lowest accuracy is around 60% for the six-class case in both datasets. These results also show that the higher accuracy will be found when the precision was lower because the accuracy of all four methods will drop from about 90% to around 60% with an increase in the number of energy class labels.

From the figures we see that the Naïve Bayes Classifier performs worse than the other three classifiers. This is because it is based on the simplified assumption that the feature values are conditionally independent given the target value. On the contrary, the features that we use, are not conditionally independent. For example, the motion sensors associated with an activity is used to find the total number of times motion sensor events were triggered and also the kinds of motion sensors involved.

To analyze the effectiveness of decision tree feature selection, we apply the ANN algorithm to both datasets with and without feature selection. From Figure 8, we can see the time efficiency has been improved greatly using feature selection. The time for building the training model drops from around 13 seconds to 4 seconds after selecting the features with high information gain. However, as seen in Figure 9, the classification accuracy is almost the same or a slight better than the performance without feature selection. The use of feature selection can improve the time performance without reducing the accuracy performance in the original data set.

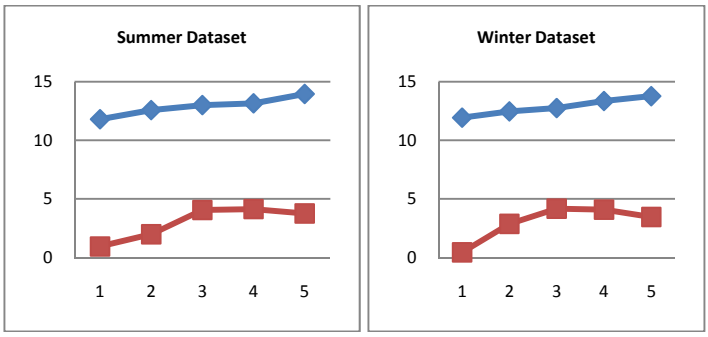


Figure 8. Comparison of time efficiency. (1:2-class; 2:3-class; 3:4-class; 4:5-class; 5:6-class; Y-axis: second; Red: with feature selection; Blue: without feature selection). Time is plotted in seconds.

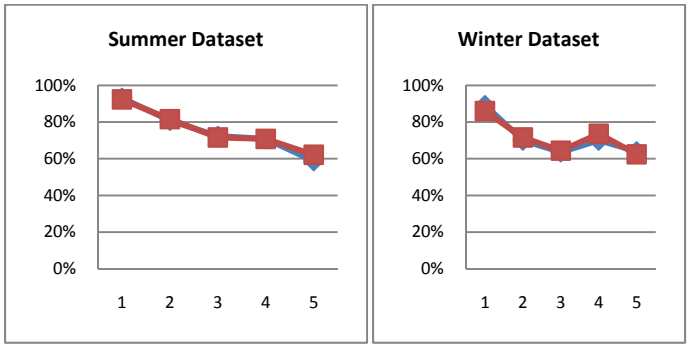


Figure 9. Comparison of prediction accuracy. (1:2-class; 2:3-class; 3:4-class; 4:5-class; 5:6-class; Red: with feature selection; Blue: without feature selection).

Figure 10 compares the performance of the ANN applied to the winter and summer data sets. From the graph, we see that the performance for the summer data set is shade better than the performance for the winter dataset. This is likely due to the fact that the room and floor heater appliances are used during the winter season, which consume a large amount of energy and are less predictable than the control of other electrical devices in the apartment.

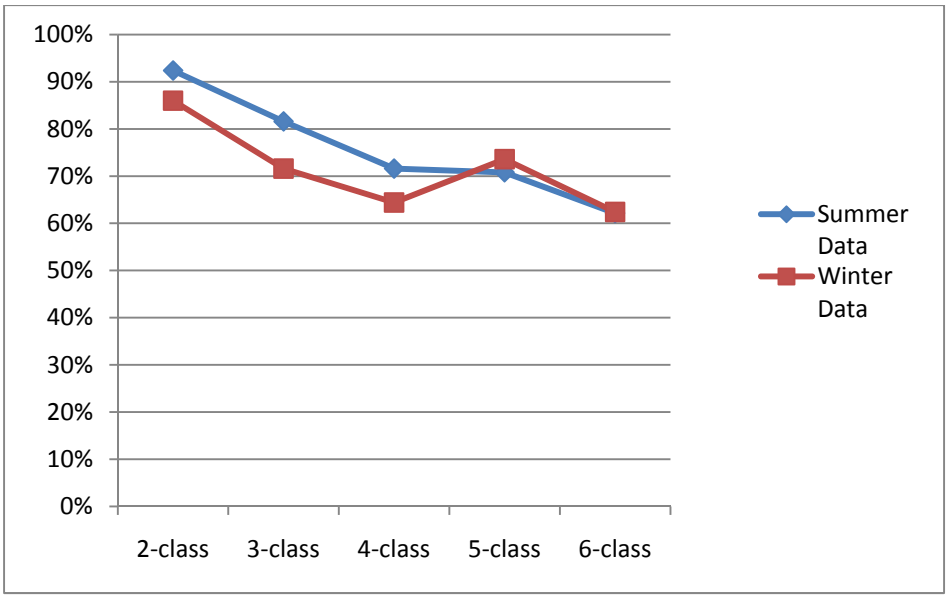


Figure 10. Comparison of the accuracy between two datasets.

Analyzing these results, we see that machine learning methods can be used as a tool to predict energy usage in smart home environments based on the human's activity and mobility. However, the accuracy of these methods is not as high as we anticipated when the energy data is divided into more than three classes. There are several reasons that lead to low performance of these algorithms. One reason is that some of the major devices are difficult to monitor and predict, such as the floor heater, which may rely on the outdoor temperature of the house. Another reason is that there is no obvious cycle of people's activities. An additional factor we can't ignore is that there is some noise and perturbation motion when the sensors record data and transfer them into the database. Finally, the sensor data we collect is not enough to predict energy precisely. As a result, we intend to collect more kinds of sensor data to improve the prediction performance.

## ENERGY TREND AND ANOMALY DETECTION

To achieve the second step of our goal, we employ statistical methods to analyze trends and look for anomalies in energy data that is collected in the CASAS testbeds. In this chapter, 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, $\bar{x}$ chart, and CUSUM chart. We test these three methods on energy data collected in the CASAS smart home apartment testbed.

The box plot [15] is a quick graphic approach for examining one or more sets of data. A box plot usually displays five important parameters describing a set of numeric data: 1) lowest value, 2) lower quartile, 3) median, 4) upper quartile, and 5) highest value. As shown in Figure 11, 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 lowest and highest values 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 help indicate the degree of spread and skewness in the data set.

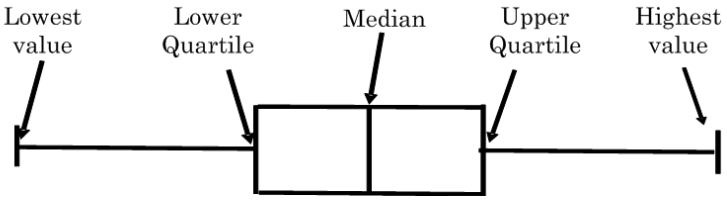


Figure 11. Configuration of a box plot.

In this study, we make use of the box plot to identify the outliers in the collected energy data, which represent those periods of time where the energy consumption lies unusually far from the main body of the data. Because even a single outlier can drastically affect the values of the mean, $\bar{x}$ and the sample deviation, $s$, a box plot is based on measures that could be resistant to the presence of the outliers. A measure of spread that is resistant to the outliers is the inter-quartile range or IQ, calculated as $IQ = UpperQuartile - LowerQuartile$. Any sample data farther than *1.5*IQ* from the closest quartile is an outlier. An outlier is extreme if it is more than *3*IQ* from the nearest quartile and it is mild otherwise.

Statistical Process Control (SPC) [16] is the application of statistical charting techniques for detecting shifts in mean or variability of a process. While SPC is applied most commonly to controlling a product's quality, it encompasses a much broader scope of applications including: data and process analysis, experimental design and decision making. 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. 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 $\bar{X}$ denotes the sample mean for a random sample of size *n* selected at a particular time, the $\bar{x}$ chart for determining control limits first calculates the mean $E(\bar{X}) = \mu$ and standard deviation $\sigma_{\bar{X}} = \sigma / \sqrt{n}$ of the sample values. Next, upper and lower control limits are defined as $\{\mu + 3\sigma_{\bar{X}} / \sqrt{n}, \mu - 3\sigma_{\bar{X}} / \sqrt{n},\}$. 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 $\bar{x}$ chart can identified the location of this change.

The disadvantage of a $\bar{x}$ 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. Cumulative sum (CUSUM) control charts [17] 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 3.

$$S_n = \sum_{i=1}^{n} (\bar{x}_i - \mu_0) \tag{3}$$

As shown in Figure 12, 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 if any of the plotted points lies outside the V-mask, either above the upper arm or below the lower arm. In the graph of Figure 12, 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.

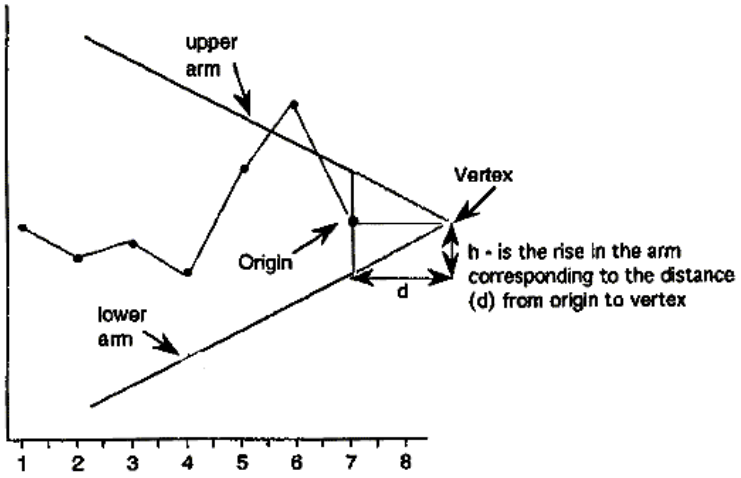


Figure 12. A sample V-Mask demonstrating an out-of-control process.

## Experimental Results

We performed two series of 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 $\bar{x}$ 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 [18], 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. As shown in Figures 13 to 15, 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 $\bar{x}$ charts for detecting energy data outliers.

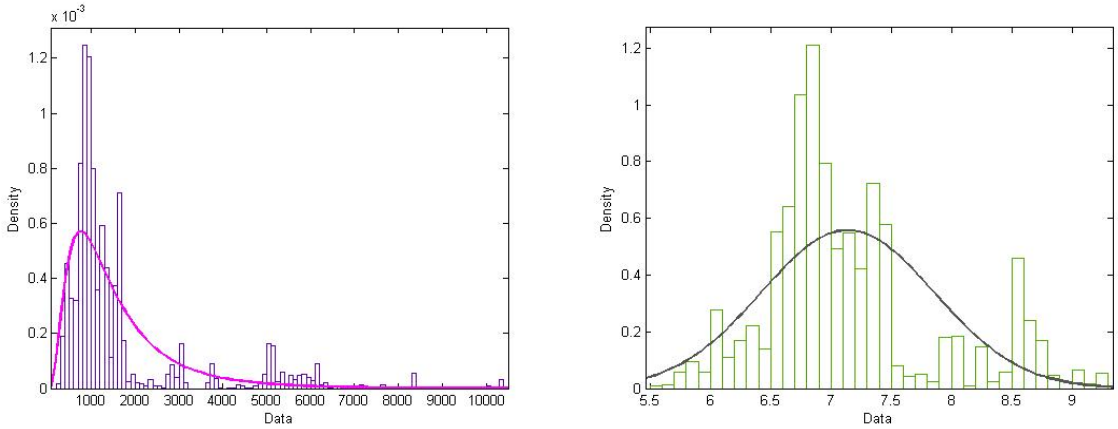


Figure 13. The lognormal and normal distribution of energy data (W) collected in the CASAS testbed over the course of one day.

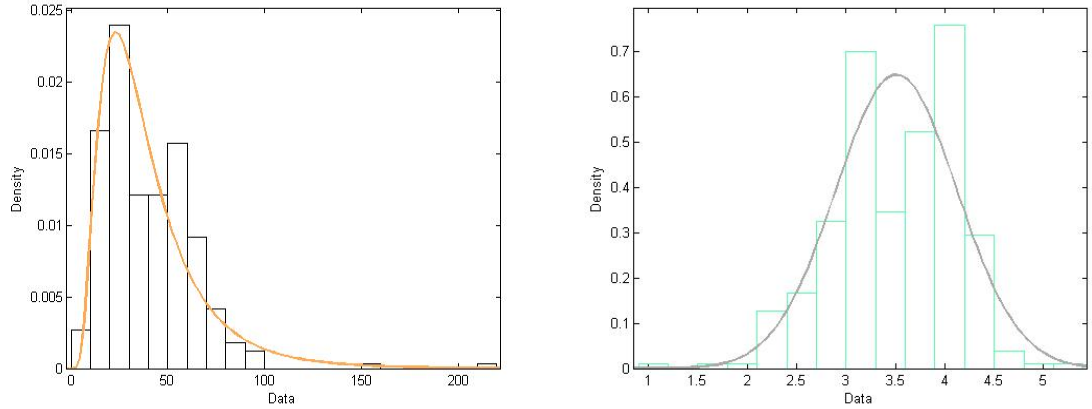


Figure 14. The lognormal and normal distribution of energy data (Kwh) for one day.

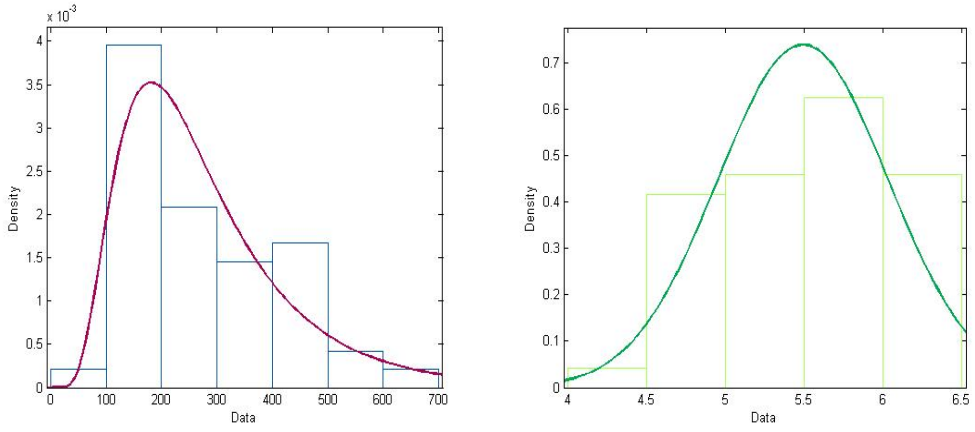


Figure 15. The lognormal and normal distribution of energy data (Kwh) for one week.

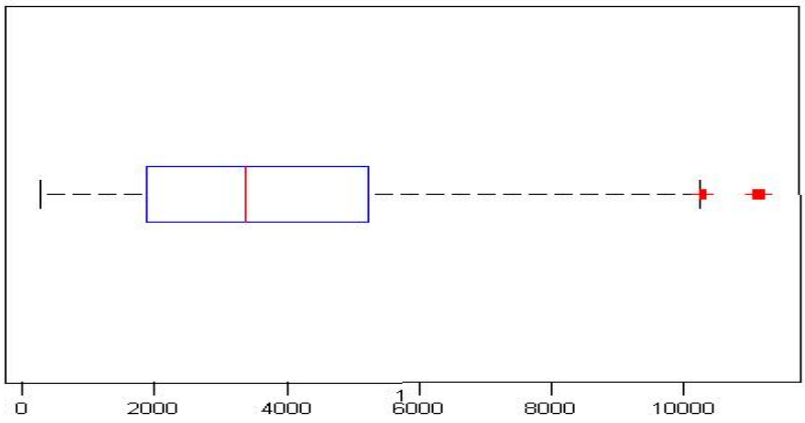


Figure 16. Box plot of wattage energy data for one day.

For the first experiment we focus on energy wattage data collected 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 16 shows the box plot graph of the data. The red points located on the right side represent the outliers. We examined those outliers in detail and found out these abnormal data occur during two main time intervals. The first set of anomalies were 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 the 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.

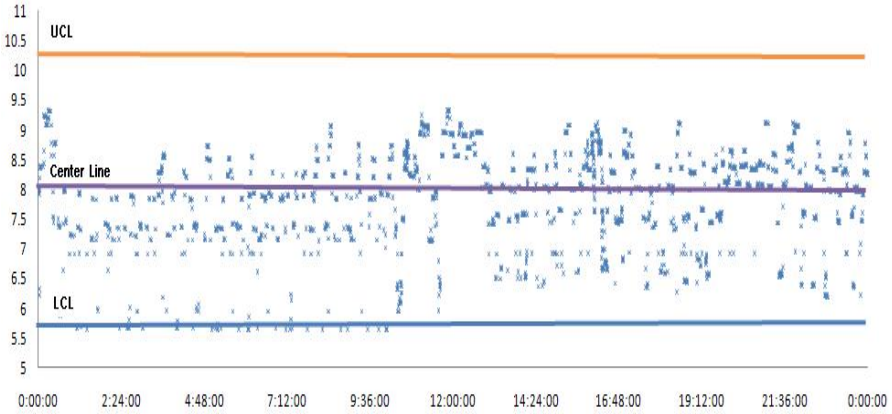


Figure 17. A $\bar{x}$ chart of energy wattage data for one day.

For the $\bar{x}$ control chart shown in Figure 17, 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 and most of the appliances are idle during that time interval.

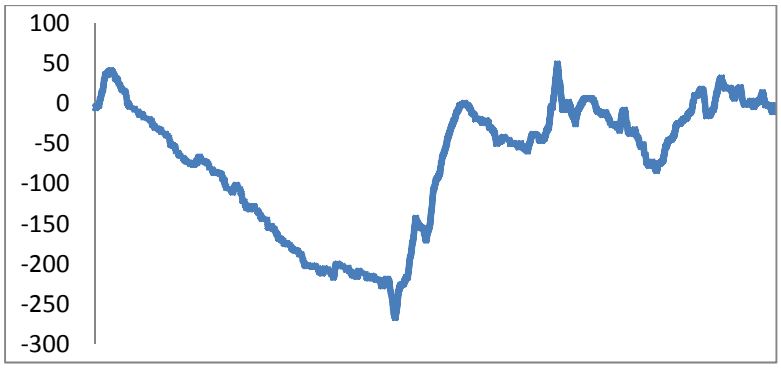


Figure 18. A CUSUM chart of energy wattage data for one day ($\theta = 0.3, h = 30$).

The CUSUM chart as described in Figure 18 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, the 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, many of which are difficult to explain. However, some of these results provide some valuable information for understanding human behavior in the smart apartment. Novelties were detected at times 00:31:07 and 16:12:50, which both represent turning points in energy usage during the day. After these times, energy consumption decreased consistently, perhaps because some large electrical devices were turned off.

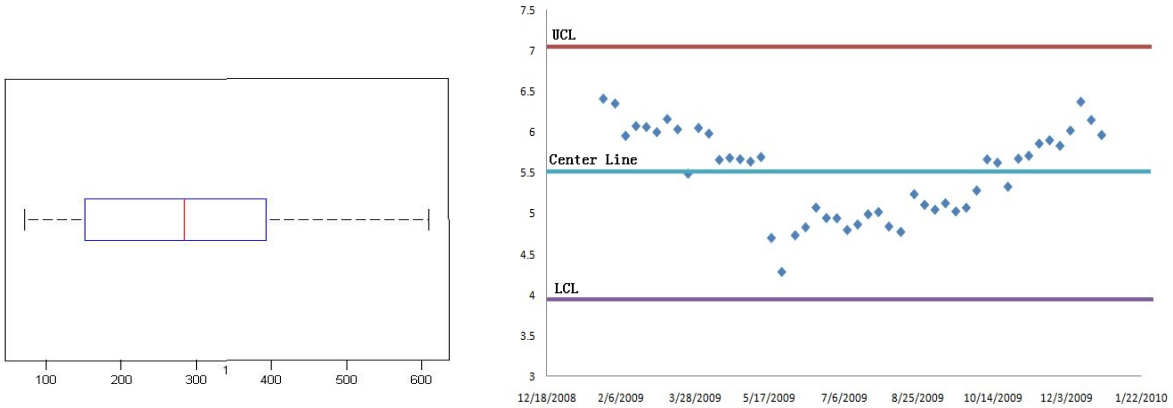


Figure 19. Bot plot chart (left) and $\bar{x}$ chart (right) of energy data (in Kwh) by week for one year.

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 19, none of the outliers can be detected by box plots or $\bar{x}$ control charts. Thus, the variation of this data set experience extreme changes during the year.

As shown in Figure 20, the CUSUM chart shows the periodic pattern of the cumulative sum. The CUSUM chart identifies four energy data abnomalies (on the dates 03/16, 05/25, 08/13, and 12/12), which represent the turning points of four different seasons (Spring, Summer, August and Winter). That result gives us a cue that there may be a possible strong relationship between seasonal temperature changes and energy usage. Thus, we continue to explore this relationship as shown in Figure 21. Figure 21 describes the change trend of the energy usage and the average temperature. Historic average regional temperature values are obtained online [19]. This figure demonstrates 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 also increase or decrease correspondingly. The likeliest reason is that the heaters in our smart home environment will consume different amounts of energy with temperature changes. In our testbed, the heaters are key influences on energy efficiency. In the future, residents might be able to utilize other heating sources such as open blinds or decrease temperature at night in order to improve energy efficiency in the apartment.

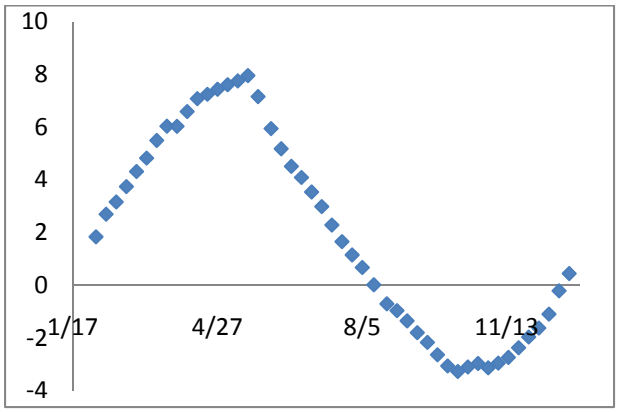


Figure 20. CUSUM chart of energy data (Kwh) by week during one year ( $\theta = 0.3, h = 0.95$ ).

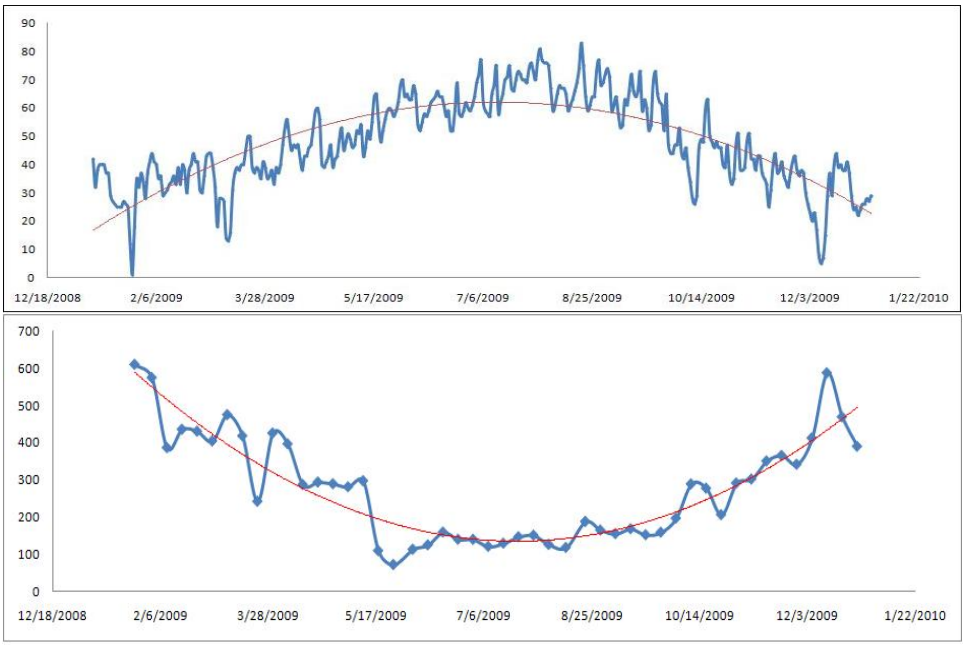


Figure 21. The comparison between temperature (top) and energy usage (bottom).

Analyzing the results of our experiments, we see that statistical approaches can be used useful tool for detecting and identifying anomalies in energy usage, which in turn provides insights on human behavior 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 $\overline{x}$ charts only detect relatively big 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 past history of the process. However, CUSUM charts are relatively slow to respond to large changes.

## CONCLUSIONS

In this chapter, we introduce several techniques for predicting energy usage and for detecting novelties, or anomalies, in energy usage which can provide insight on human behavior. To assess the performance of our algorithms we provided experimental results using real data collected in the CASAS smart environment testbeds.

In our ongoing work, we plan to investigate methods to detect a greater range of anomalies. We also plan to install more sensitive power meters in order to capture more accurate changes in energy consumption. Our future plans also include collecting data in a greater variety of households, which will allow us to determine whether energy predictions, energy usage trends, and energy anomalies exist and generalize across multiple settings.

## REFERENCES


Bauer, G., Stockinger, K., & Lukowicz, P. (2009). Recognizing the Use-Mode of Kitchen Appliances from Their Current Consumption. *Smart Sensing and Context*, 163–176.

Bellman, R. E. (1961). *Adaptive control processes - A guided tour*. Princeton, New Jersey, U.S.A.: Princeton University Press.

Beware. (2010), Retrieved May 25, 2010, from http://www.energyawareness.eu/beaware

Boser, B. E., Guyon, I. M., & Vapnik, V. N. (1992). A training algorithm for optimal margin classifiers. *Proceedings of the fifth annual workshop on Computational learning theory* (pp. 144–152).

Chen, C., Das, B., & Cook, D. J. (2010). *Energy prediction based on resident's activity*. *Proceedings of the International workshop on Knowledge Discovery from Sensor Data*.

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

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

Devore, J. L. (2008). *Probability and Statistics for Engineering and the Sciences*: Cengage Learning.

Energy Efficiency in Buildings. (2009), Retrievecd October 11, 2009, from www.wbcsd.org

Gao, J., Hauptmann, A. G., Bharucha, A., & Wactlar, H. D. (2004). Dining activity analysis using a hidden markov model *Proceedings of the 17th International Conference on Pattern Recognition.*

Hu, H., Pan, J., Zheng, W., Liu, N., & Yang, Q. (2008). Real world activity recognition with multiple goals *Proceedings of the 10th international Conference on Ubiquitous computing* (pp. 30–39).

Krumm, J., Harris, S., Meyers, B., Brumitt, B., Hale, M., & Shafer, S. (2000). Multi-Camera Multi-Person Tracking for EasyLiving *VS '00: Proceedings of the Third IEEE International Workshop on Visual Surveillance* (pp. 3). Washington, DC, USA.

Liu, H., Hussain, F., Tan, C. L., & Dash, M. (2002). Discretization: An enabling technique. *Data Mining and Knowledge Discovery, 6*(4), 393–423.

Mitchell, T. (1997). *Machine Learning*. New York: AMcGraw Hill.

Mozer, M. C. (1998). The neural network house: An environment that adapts to its inhabitants *Proceedings of the American Association for Artificial Intelligence Spring Symposium on Intelligent Environments* (pp. 110–114).

Orr, R. J., & Abowd, G. D. (2000). The smart floor: A mechanism for natural user identification and tracking *Proceedings of the Conference on Human Factors in Computing Systems* (pp. 275–276).

Patel, S. N., Robertson, T., Kientz, J. A., Reynolds, M. S., & Abowd, G. D. (2007). At the flick of a switch: Detecting and classifying unique electrical events on the residential power line. *Proceedings of the International Conference on Ubiquitous Computing* (pp. 271).

Patel, S. N., Truong, K. N., & Abowd, G. D. (2006). PowerLine Positioning: A Practical Sub-Room-Level Indoor Location System for Domestic *Ubiquitous Computing* 4206;441-458, Springer Berlin / Heidelberg.

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

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

Rice, J. (1995). Mathematical Statistics and Data Analysis. *Duxbury Press*.

Rish, I. (2001). An empirical study of the naive Bayes classifier *Proceedings of the IJCAI Workshop on Empirical Methods in Artificial Intelligence* (pp. 41–46).

Sánchez, D., Tentori, M., & Favela, J. (2008). Activity recognition for the smart hospital. *IEEE Intelligent Systems, 23*(2), 50–57.

Singla, G., Cook, D. J., & Schmitter-Edgecombe, M. (2010). Recognizing independent and joint activities among multiple residents in smart environments. *Journal of Ambient Intelligence and Humanized Computing*, 1–7.

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

Tapia, E., Intille, S., Lopez, L., & Larson, K. (2006). The design of a portable kit of wireless sensors for naturalistic data collection. *Pervasive Computing*, 117–134.

Tapia, E. M., Intille, S. S., & Larson, K. (2004). Activity recognition in the home using simple and ubiquitous sensors. *Pervasive Computing*, 158–175.

Thompson, J. R., & Koronacki, J. (2002). *Statistical process control: the Deming paradigm and beyond*: CRC Pr I Llc.

Tukey, J. W. (1977). *Exploratory data analysis*. Reading, MA: Addison-Wesley.

Weather Underground. (2010), Retrieved June 26, 2010, from www.wunderground.com

Witten, I. H., & Frank, E. (2005). *Data Mining: Practical machine learning tools and techniques*: Morgan Kaufmann Pub.

Zheng, V. W., Hu, D. H., & Yang, Q. (2009). Cross-domain activity recognition *Proceedings of the International Conference on Ubiquitous Computing* (pp. 61–70).

Zornetzer, S. F. (1995). *An introduction to neural and electronic networks*: Morgan Kaufmann.