# Energy Outlier Detection in Smart Environments

**Chao Chen and Diane Cook**
School of Electrical Engineering and Computer
Washington State University
Pullman, USA
{cchen, cook}@eecs.wsu.edu

### Abstract

Despite a dramatic growth of power consumption in households, less attention has been paid to monitoring, analyzing and predicting energy usage. In this paper, we propose a framework to mine raw energy data by transforming time series energy data into a symbol sequence, and then extend a suffix tree data structure as an efficient representation to analyze global structural patterns. Then, we use a clustering algorithm to detect energy pattern outliers which are far from their cluster centroids. To validate our approach, we use real power data collected from a smart apartment testbed during two months.

## Introduction

In smart environment research, most of the effort has been directed towards activity recognition with applications in health monitoring. Energy consumption is an aspect of home life that is often overlooked. This oversight is detrimental. Between 1973 and 2004, energy consumption increased at a higher rate than the population growth (Pérez-Lombard, Ortiz, and Pout 2008). This growth is not entirely due to manufacturing plants and automobiles. In fact, households are responsible for over 40% of energy usage in most countries (Pérez-Lombard, Ortiz, and Pout 2008). As a result, there is an urgent need to develop technologies that examine energy usage in homes and to encourage energy efficient behaviors, in addition to energy efficient devices in households.

Earlier studies have shown that home residents reduce energy expenditure by 5-15% on average just as a response to acquiring and viewing raw usage data (Darby 2006). Traditional power meters provide only basic consumption data such as current power usage and killowatt hour. There is a clear need for improving householders' working knowledge of their behaviors and energy consumption. Pervasive computing techniques can improve the quality of information supplied to users by identifying usage trends and anomalies, and providing users with suggestions about how to save energy and conserve natural resources.

We hypothesize that providing users with behavior-based knowledge of energy consumption, suggestions for energy reduction, and automation support will result in more substantial decreases in overall consumption. This view is supported by an increasing body of work that links awareness of energy consumption and its impact to behavioral change (Darby 2010). In our work we propose to use smart homes and pervasive computing techniques to provide these important insights.The long-term vision for this project is to enhance understanding of human resource consumption and to provide resource efficiency in smart homes. We hypothesize that patterns and anomalies can be automatically detected from energy consumption data and that these discoveries can provide insights on behavioral patterns. The proposed system preprocesses power data into a symbol sequence, then discovers energy patterns using a suffix tree (Gusfield 1997). The results of this work can be used to give residents feedback on energy consumption and also be used to remove some erroneous energy records in the database.

This paper is organized as follows: Section 2 discusses related work. Section 3 introduces our system architecture. Sections 4-6 describe different modules in the system. Section 7 presents the results of our experiments. A discussion of the current approach and future work is given in Section 8.

## Related Work

In this paper, we mainly focus on domains of energy research. There are two domains of energy researches that are related to our work: 1) non-intrusive appliance load monitoring, and 2) energy conservation services.

A non-intrusive appliance load monitor (Hart 1992) is designed to detect the turning on and off of individual appliances in a electrical circuit. A few studies have focused specifically on non-intrusive appliance detection. Kato et al. (Kato et al. 2009) extracted features from power waveform by Linear Discriminant Analysis (LDA) and used support vector machines (SVM) to classify appliances. Gupta et al. (Gupta, Reynolds, and Patel 2010) analyze frequency electromagnetic interference (EMI) on the power line, and then uses SVM to identify unique occurrences of switching events.

With respect to energy conservation, some works focus on providing energy information service and saving tips to the residents. Google PowerMeter (Google 2010) is a free energy monitoring tool for saving energy and money by pro-

viding energy information via smart meters and energy monitoring devices. Microsoft Hohm (Microsoft 2010) is a web service that can predict the energy distribution of the house and suggest suitable energy saving tips to user. Our previous work (Chen, Das, and Cook 2010) predicted energy consumption based on sensor data collected and generated by the residents in a smart home environment.

Table 1: Transaction Data vs. Energy Data

| Transaction ID | Items |
|---|---|
| 1 | {A,B,C,D,F,G,H} |
| 2 | {D,F,G,H} |

| Timestamp | Power Value (Wattage) |
|---|---|
| 2009-06-02 00:00:32 | 930 |
| 2009-06-02 00:00:38 | 471 |

Many approaches have been proposed for discovering sequential patterns in data. However, most of these methods only consider source data to be in a transactional format. In our smart environment, the energy data is numeric and arrives in a continual stream. As shown in Table 1, each transaction is identified by a unique transaction ID associated with a set of items. In contrast, an energy reading is composed of two attributes: a timestamp and a numeric value, which is a continuous stream of sensor events over time. To mine energy sensor data, we use the equal-width binning method (Liu et al. 2002) to discretize the energy readings into the symbol sequence, which can be applied as a sequence pattern mining method. To the best of our knowledge, this is the first work applying pattern discovering approach into detecting energy outliers in the home environments.

## System Architecture

In this paper, we developed a prototype system framework for energy data collection, energy data transformation and energy pattern outlier detection as shown in Figure 1.

- The smart environment contains sensors, controllable devices, and software.
- Energy data transformation is a preprocessing layer which analyzes and transforms raw energy data to a symbol string sequence using a binning approach. A Suffix Tree Generator is responsible for generating energy patterns after inputting a given energy symbol sequence.
- Energy pattern outlier detection is an algorithm which implements energy outlier detection. The output from this module can provide feedback to the users which can inform their choices and improve their energy consumption.

## Smart Home Environment

The smart home environment testbed that we are using to analyze energy usage is a three bedroom apartment. Two volunteer participants regularly live in the testbed. Thus, the data we collected are from the real life of the participants.

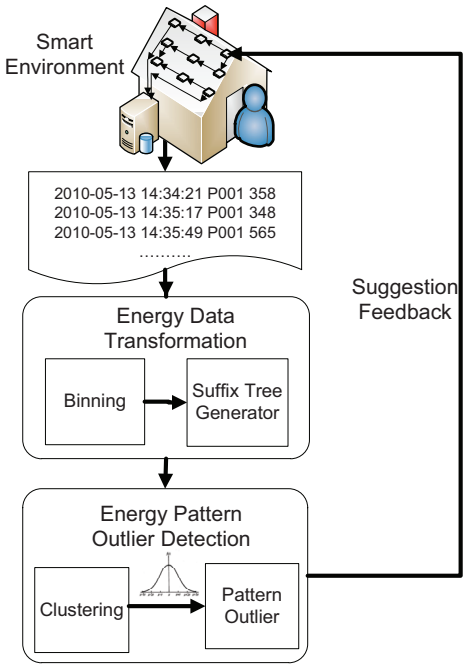


Figure 1: System architecture of our energy outlier detection tool.

To track residents' mobility, we use motion sensors placed on the ceilings as shown in Figure 2. A simple power meter records the amount of instantaneous power usage at a fixed sampling frequency and the total amount of power which is used over time. An in-house sensor network captures all sensor data and stores it in a SQL database. The sensor events annotated with the corresponding activities being performed while the sensor events were generated.

## Energy Data Transformation

The important step in utilizing smart home technologies for energy efficiency is to analyze normal patterns of usage and identify abnormal, or anomalous situations. 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.

To begin, a formal definition of the dataset is required.

**Definition 1.** *Let $e = (t, v)$ be an individual **energy senosr event** in our smart environment, where $t$ refers to the timestamp when $v$ has been activated, and $v$ refers to an energy numerical value.*

For data mining purpose, we are typically not interested in

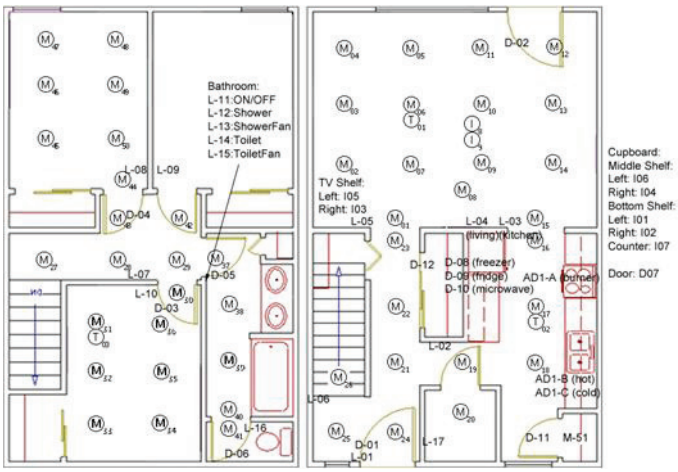


Figure 2: Three-bedroom smart apartment used for our data collection (motion (M), temperature (T), water (W), burner (B), telephone (P), and item (I)).

any individual energy sensor event; rather, we are interested in a global energy sensor sequence:

**Definition 2.** *An **energy sensor sequence** $E = e_1 e_2 \dots e_n$ is an ordered set of n energy sensor events.*

Smart home power meters record the amount of instantaneous power that is currently being consumed in real time. We first discretize this data into $k$ value ranges using equal-width binning and then convert the value ranges to symbols. This representation allows symbolic approaches to be applied to analyzing the data, at the risk of losing some precision in the values. Through binning, an energy sensor sequence $E$ can be transformed into a new energy symbol sequence S, which is defined as:

**Definition 3.** *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 which employs suffix trees (Gusfield 1997). Unlike other data mining methods, which are exponential in their complexity, this algorithm can contribute 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 4.** *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 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.

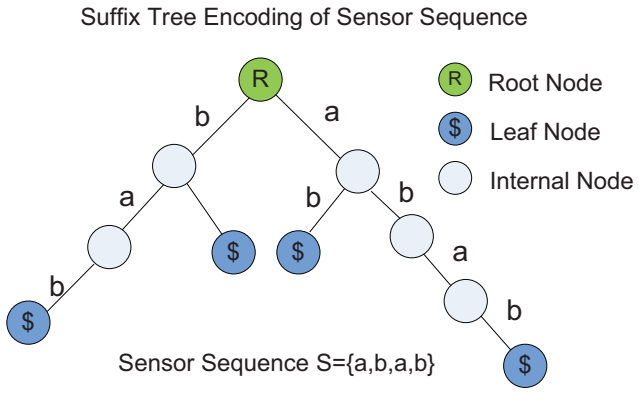


Figure 3: A suffix tree defined on a symbol sequence $S$ with length $m$ can represent every subsequence in $S$ with at most $2m$ nodes

A graphical illustration of the transformation of an energy sequence into its equivalent suffix tree is shown in Figure 3. 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 5.** *Let an **energy pattern** $p_i$ in $S$ represent the subsequence generated by 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 2 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.

## Energy Pattern Outlier Detection

In this section, we assume all energy patterns with the same length are similar to each other, which will fall into a unique cluster. To detect abnormal situations, we then cluster all the

Table 2: Examples of Energy Pattern

| Energy Pattern | Pattern Length | Raw Energy | Pattern Frequency |
|---|---|---|---|
| CC | 2 | 752<br>742 | 26952 |
| ZFZ | 3 | 5000<br>1021.2<br>5007 | 13 |

energy patterns with the same length using Euclidean distance, as described in the next section. 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 (Fisher 1987), 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 Euclidean distance measure, which is a geometric distance in the multidimensional 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 distance of 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 alphanumeric 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:

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

**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:

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

**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:

$$d_3(p^1, p^2) = |f(p^1) - f(p^2)| \tag{3}$$

It should be noted that the frequency of energy pattern outliers mainly divided into two parts: 1) low frequency, 2) high frequency. In this paper, we are more interested in the low-frequency energy outliers since they have a high probability of being outliers.

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:

$$d(p^1, p^2) = \sqrt{d_1(p^1, p^2)^2 + d_2(p^1, p^2)^2 + d_3(p^1, p^2)^2} \tag{4}$$

In the second step of our analysis, we use the defined 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 $c_p$. In the second stage, we set an outlying factor $o_{p^k}$ for each energy pattern $p^k$ in the cluster. This factor depends on its distance from the centroid of the cluster. We define the outlying factor of an energy pattern $p^k$ in the cluster as follows:

$$o_{p^k} = d(p^k, c_p) \tag{5}$$

The greater the outlying factor is, the more likely it is the pattern is an outlier. As shown in Figure 4, the energy patterns for which $o_{p^k} > threshold$, are defined as the outliers.

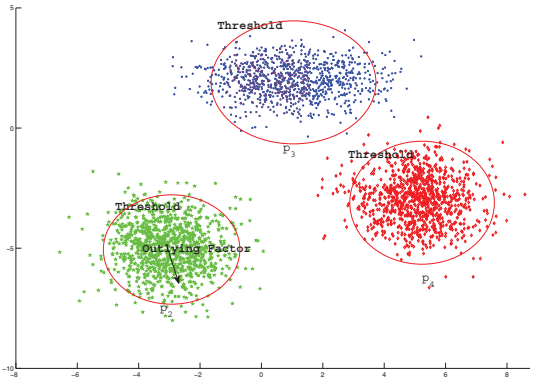


Figure 4: Example of outlier detection in the cluster.

From this discussion it is apparent that the choice of a threshold value greatly influences the selection of outliers. To determine the value for this application domain we plot a histogram of all pattern distance values to the centorid (also referred to as outlying factors, see Figure 5). It was

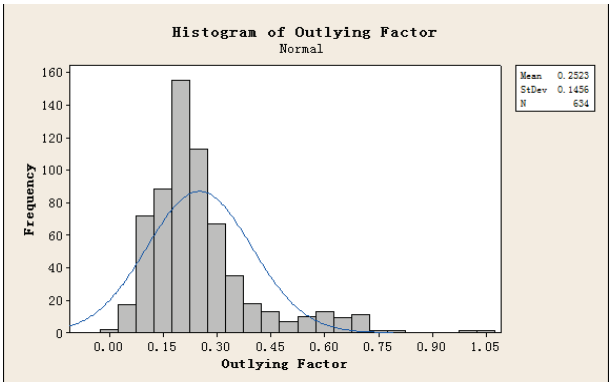


Figure 5: Histogram of outlying factors of all $p_2$ energy patterns ($k = 50$).

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.

To provide a basis of comparison, an outlier detection was performed on the energy usage data using a standard box plot analysis (Tukey 1977). The box plot is a quick graphic approach for examining 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, the box plot is constructed by drawing a rectangle between the upper and lower quartiles with a solid line indicating the median. The lower and upper fences exist at the boundary of the solid line.

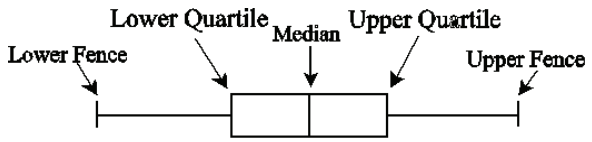


Figure 6: Configuration of a box plot.

In our study, we use the box plot as an alternative method to identify 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. If $Q_1$ and $Q_3$ are the lower quartile and the upper quartile, a measure of spread that is resistant to the outliers is the inter-quartile range or $IQ$, calculated as $IQ = Q_3 - Q_1$. As shown in Figure 6, the fences lie at $Q_1 k * IQ$ and $Q_3 + k * IQ$. The change of the value of $k$ can affect the number of the observations outside the fence. For this work, a value of $k = 1.5$ was used, which has been indicated as acceptable for most situations (Frigge, Hoaglin, and Iglewicz 1989). 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.

## Experiment Result

The outlier detection experiment we performed uses the sensor data generated by two residents during two summer months at the smart apartment testbed.

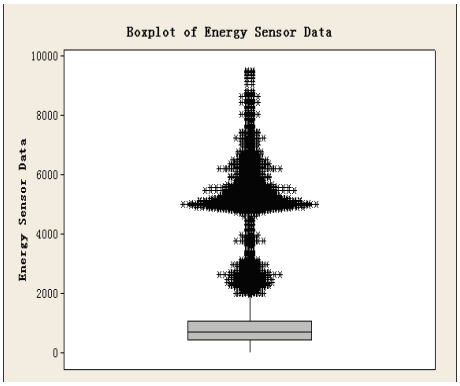


Figure 7: Energy outlier detection using the boxplot.

For the first experiment, a total of 95,968 power events were collected. Figure 7 shows the result of the boxplot approach on this dataset. The black points located on the top represent the outliers. The boxplot considered 12,718 sensor events as potential outliers, since it is merely able to detect energy consumption that lies unusually far from the main body of the data. However, it is difficult for most users to determine which outliers are true outliers and identify potential reasons for these outliers, because there are too many false positives.

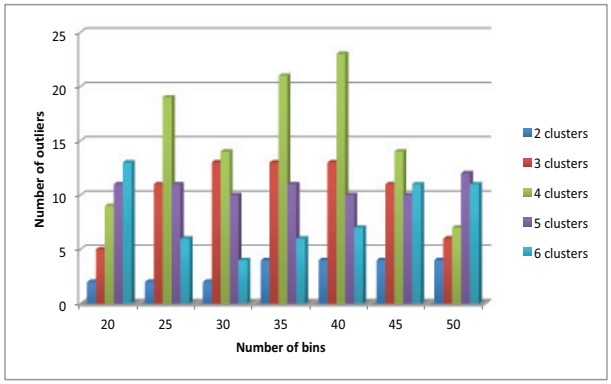


Figure 8: Distribution of number of pattern outliers using our clustering approach.

Next, we use our proposed clustering algorithm to analyze the same power dataset. Figure 8 depicts the distribution of energy patterns that were detected as potential outliers under different numbers of bins and clusters. Comparing our method with the boxplot, it shows that the number of the outliers reported by the clustering approach has been decreased notably. This increases the chance to accurately determine real outliers in the dataset.

Table 3 displays the result of our clustering method with several different pattern lengths when $k$ is assigned to 50.

Table 3: Experimental Results of Outlier Detection ($k = 50$)

| Pattern Length | Number of Outliers |
|---|---|
| 2 | 24 |
| 3 | 54 |
| 4 | 25 |
| 5 | 23 |
| 6 | 34 |

To explore potential reasons for the anomalous usage patterns, those outliers were examined in detail. It was discovered that these abnormal events occur in two main ways. The first set of outliers was mainly due to large changes in energy usage, or 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 the others and can create anomalies when there are long showers. In addition, during the middle of the day is the residents cooking time 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 dramatical increase in energy consumption. To response to these outliers, the residents can analyze their energy needs during these activities to identify energy-saving behaviors.

Table 4: Example of an Outlier.

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

The outliers in the second set were found to be two successive energy events, whose values are different but occur at the same time, as shown in Table 4. 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 removed and subsequently improve the accuracy of additional analysis methods. Therefore, we checked the entire dataset for these types of outliers. The result was that 6,398 entries from Kyoto 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 centroid of the clustering, the rate of false negative may be relatively higher, which means that lots of real outliers are likely to be ignored by this approach. One possible solution is to decrease the predefined threshold, which makes our approach to detect more outliers at the risk of increasing the rate of false positive.

## Conclusions

In this paper, we introduce a data mining and clustering technique for detecting the outliers and anomalies in energy usage. We first use an equal-width binning approach to translate raw energy data into a symbol sequence, and then extend a suffix tree to generate energy patterns. Through clustering these energy patterns, we detect the energy outliers which are far from their cluster centroid. The purpose of outlier detection is to find some extreme energy change power, which may lead to potential security problems in the smart environment.

In our ongoing work, we plan to investigate methods to detect a greater range of anomalies. Additionally, we also plan to install more sensitive power meters in order to capture more changes and patterns 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


Chen, C.; Das, B.; and Cook, D. J. 2010. Energy prediction based on resident's activity. In *the International Workshop on Knowledge Discovery from Sensor Data*.

Darby, S. 2006. The effectiveness of feedback on energy consumption. Online.

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

Fisher, D. H. 1987. Knowledge acquisition via incremental conceptual clustering. *Machine Learning* 2(2):139–172.

Frigge, M.; Hoaglin, D.; and Iglewicz, B. 1989. Some implementations of the boxplot. *The American Statistician* 43(1):50–54.

Google. 2010. Google Power Meter. http://www.google.com/powermeter/about/.

Gupta, S.; Reynolds, M. S.; and Patel, S. N. 2010. 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*, 139–148.

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

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

Kato, T.; Cho, H.; Lee, D.; Toyomura, T.; and Yamazaki, T. 2009. 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* 150–157.

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

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

Microsoft. 2010. Microsoft Hohm. http://www.microsoft-hohm.com/.

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

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