Application of Clustering Techniques to Energy Data to Enhance Analysts' Productivity

Wendy Foslien, Honeywell Labs
Valerie Guralnik, Honeywell Labs
Steve Harp, Honeywell Labs
William Koran, Honeywell Atrium

ABSTRACT

Honeywell Atrium™ collects and analyzes its customers' utility data to study building operations, energy usage, equipment performance and diagnostics for the purpose of reducing customers' operational costs. Each Atrium™ analyst is responsible for supporting several hundred buildings and supporting customers with recommendations for cost reduction. This makes it necessary to develop tools that will direct analysts towards problem facilities, rather then manually scanning all available data.

In this paper we present two different applications of clustering analysis to discover anomalous data. The first application identifies facilities that operate with unusual electrical load profiles, based on characteristics of multiple facilities' electric operational data. The second application identifies days exhibiting abnormal operation within individual facilities. Our experiments using data sets derived from energy consumption profiles of a major retailer show that clustering analysis is a useful tool in guiding analysts to identify anomalous buildings and daily profiles and enhancing analysts' productivity.

Introduction

In many businesses today, large volumes of time series data are collected and stored in databases. Making good use of this data is challenging, most often because detailed and time consuming analysis is not feasible. This paper addresses the automation of time series data analysis, for the building energy domain. We use a clustering approach to make decisions about how to group related segments of data. The attributes of the clusters are then used to make inferences about why these segments of data are grouped together.

One example of a large repository of time series data is found in the Honeywell Atrium™ databases. Atrium™ is an energy information service marketed to store chains, to improve energy management across an enterprise. The service provides information with the appropriate detail for decision-makers in each organization. Energy use data is stored at 15-minute intervals for reporting and analysis, and the data is analyzed by custom applications to provide input for a group of analysts with a background in energy management. To obtain the data, the service can install meters and gateways, make use of a customer's existing metering infrastructure, or take data from third parties.

Atrium™ analysts all have a background in facility management, energy retrofit, building equipment, and building energy analysis. They have more experience in data analysis than the typical building engineer, and have access to advanced tools to help them analyze the collected data. When they identify a change that would save costs at a facility, they communicate this to the customer via a report on the customer's Atrium™ web page.

With the flexible applications and experienced engineers, the information provided by Atrium™ can be tailored to a customer's needs. Atrium™ can provide simple rankings of stores by energy use per square foot; additional comparisons by day type, time of day, occupancy, or weather conditions; track performance improvements after operational changes or retrofits; and provide recommendations for operational improvements.

The engineers review recent data to watch for anomalous energy use, and analyze long term data to understand the effect of weather or building-specific variables. After sufficient data has been collected and reviewed, the engineers recommend changes to the operation of individual buildings or changes to corporate operations that can save energy or operational costs. If additional services are required, such as more detailed audits and surveys, assistance with implementation of the recommended improvements, or storage and analysis of additional data streams, these services can be provided at an additional fee.

The data analysis process has two components: initial evaluation of the metered data and daily review of the data for unusual patterns. One purpose of the initial evaluation is to prioritize buildings for analysis. Buildings that are both high use and high cost should get the initial attention from the analyst. Another purpose of the initial analysis is to gain an understanding of typical operation, and to group buildings with similar behavior so that similarly behaving buildings can be analyzed together, again expediting the analysis. The data can also be segmented by other building attributes, such as type of lighting or type of HVAC system. These additional attributes are unique to each customer, and require custom handling, but this customization is automated to facilitate the use of these types of unique but significant data.

One of the questions that the initial analysis attempts to address is "Do all the buildings operate similarly?" If the answer is no, then the analyst needs to examine the data to understand the causes for the differences in operation. The clustering approach presented in the following sections attempts to sort the buildings into groups with similar operation, and the visualization methods used on the cluster data help the analysts use their expertise to determine the causes of variation in operation.

Daily analysis is performed by a set of data mining tools not discussed in this paper. These tools are used to direct the analysts toward buildings with the highest priority needs by identifying unusual behavior.

In the remainder of this paper, we will describe the type of cluster analysis employed, and some specifics of the selected energy data. We then show two ways to apply the analysis to find anomalous facilities or days of operation before presenting some conclusions and suggested future work.

Overview of Clustering Techniques

Clustering is the task of grouping together objects into meaningful subclasses. Agglomerative hierarchical clustering (Duda & Hart 1973) and *K*-means (Duda & Hart 1973; Kaufman & Rousseeuw 1990) are two techniques that are commonly used for clustering.

Hierarchical techniques produce a nested sequence of partitions, with a single all-inclusive cluster at the top and singleton clusters of individual objects at the bottom. Each intermediate level can be viewed as combining two clusters from the next lower level (or splitting a cluster from the next higher level). Agglomerative hierarchical algorithms start with each object as a separate cluster. After each merge, the total number of clusters decreases by one. These steps can be repeated until the desired number of clusters is obtained or the distance between two closest clusters is above a certain threshold distance.

In contrast to hierarchical techniques, partitional clustering techniques create a one-level (un-nested) partitioning of the data points. Partitional clustering attempts to break a data set into *K* clusters such that the partition optimizes a given criterion. Centroid-based approaches as typified by *K*-means try to assign objects to clusters such that mean square distance of objects to the centroid of the assigned cluster is minimized.

Data Set

The basic data set used for cluster analysis consisted of data extracted from 31 whole building meters, from the beginning of 1999 through August 2001. All analyzed buildings had a single meter. Temperature and humidity data were extracted as available, but weather data were not available for the complete set. Much of the data was imported from archives, which did not include weather data.

We also extracted data on the square footage and zip code for each building. This summary data is shown below in Table 1.

Table 1. Characteristics of Meters in Sample Set

Meter Id	Square Footage	Zip Code	Meter Id	Square Footage	Zip Code
1	106039	94040	17	59765	93277
2	118461	95121	18	66675	90630
3	105070	93612	19	78670	92781
4	105070	93306	20	75236	93065
5	82176	95037	21	116240	93030
6	116244	92860	22	124871	91801
7	99984	92543	23	112038	92840
8	115900	91773	24	122088	90815
9	115107	94550	25	105154	93534
10	1020801	94545	26	105154	92553
11	89009	94560	27	119840	91355
12	84669	95127	28	105154	91710
13	88862	95118	29	108085	92691
14	80008	94564	30	105142	93277
15	74850	95008	31	103486	93304
16	52726	94589			

Cluster Analysis

In this section we present two different applications of clustering analysis to discover anomalous data. The objective of the first application is to automatically extract facilities that operate with unusual energy load profiles, based on characteristics of multiple facilities' operational data. The second application attempts to identify abnormal daily operations of certain buildings.

Clustering Facilities

In this section we present the results of applying clustering analysis to energy consumption data of multiple buildings collected over a certain period of time. The objective is to automatically extract facilities that operate with unusual energy load profiles, based on characteristics of multiple facilities' operational data. We can think of this clustering approach as searching for buildings that have similar performance. Our expectation is that buildings that have unusual behavior (good or bad) will be grouped into smaller clusters. Note that the cluster analysis is going to tell us when buildings operate in a similar manner, but not why they operate that way.

A simple approach to clustering facilities could have been to use existing clustering algorithms with the Euclidean distance metric as a similarity measure. Unfortunately, the Euclidean distance metric is very sensitive to noise. In addition, clustering raw data would most likely not give us insight on why certain meter belongs to one or another cluster.

As a solution, we decided to use the following technique. First, we apply the Fourier transform (Hamming 1977; Oppenheim & Schafer 1975) on raw data and retain only significant transform coefficients. Then, we use the retained coefficients as features and cluster data in the feature space. There are several benefits in this approach. First, according to Parseval's theorem (Oppenheim & Schafer 1975) the Fourier transform preserves the Euclidean distance in the time or frequency domain. Second, Fourier transformation removes noise from the data set. Third, retaining only dominant coefficients reduces data dimensionality. Finally, the period of each of the retained frequencies can give analysts insight on why the meters are placed in the clusters – e.g. by examining which time period the most significant coefficients are from will indicate whether the clustering was influenced by hourly, daily, weekly or seasonal variations.

Data preparation. Some of the consumption data was missing from the original data set. Since we needed continuous data to apply cluster analysis, we filled the missing data points by using linear interpolation. Another alternative to filling data could have been to use a moving average of daily consumption in each 15-minute time segment.

Because the start date of data varied between different buildings, to make each meter's data set of equal length, we extracted data for each building covering the period from the end of August 1999 to beginning of July 2001. In the end each data set contained 65536 data points. The length of each time series was a power of two, which allowed us to use Fast Fourier transform (FFT), which is more efficient than the standard Fourier transform. Before applying the FFT, consumption was normalized by the square footage of the building to remove the size of the building as one of the factors to drive formation of the clusters. After the FFT was applied, we discovered that the dominant coefficients represented daily variations.

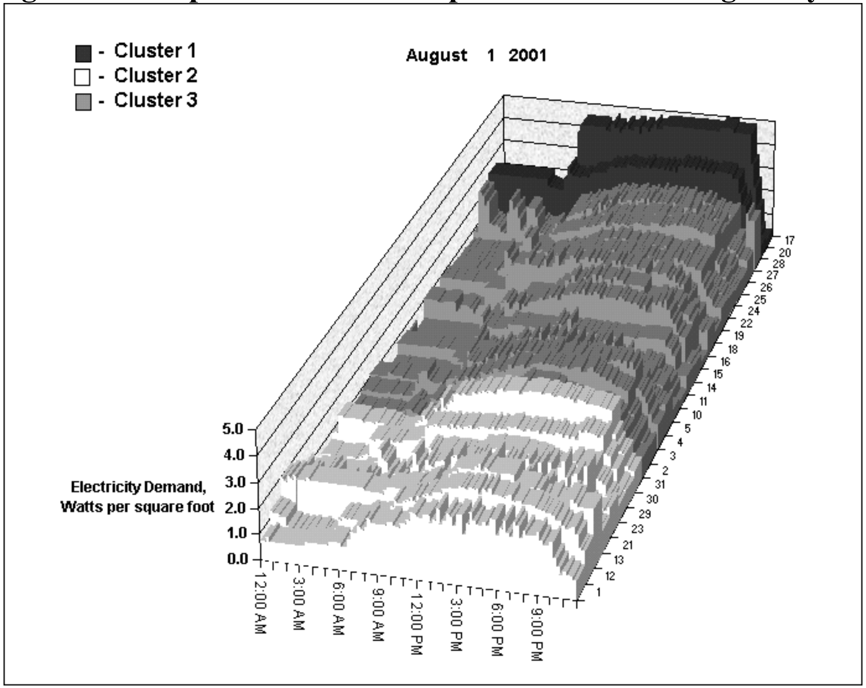
Results. Using a K-means algorithm, we clustered data 3, 4, and 5-ways, meaning that we allowed the formation of 3 groups, 4 groups, and 5 groups as separate experiments. The resulting clusters are shown in Table 2, with the meters sorted such that they fall into the groups for 3, 4, or 5 clusters. The columns labeled "3-way clustering", "4-way clustering", and "5-way clustering" show the cluster where this meter was placed in each experiment. We can see from this table that clustering results appear to be stable. Going from 3 to 4 clusters and from 4 to 5 does not cause a lot of moving of meters between clusters. Instead, increasing the number of clusters causes one large cluster to be split into two smaller ones.

Table 2. Cluster Placement for Buildings in Data Set

Meter ID	3-way clustering	4-way clustering	5-way clustering	Meter Id	3-way clustering	4-way clustering	5-way clustering
17	1	1	1	2	3	4	4
20	1	1	1	4	3	4	4
1	2	2	2	5	3	4	4
13	2	2	2	24	3	4	4
21	2	2	2	25	3	4	4
23	2	2	2	26	3	4	4
29	2	2	2	27	3	4	4
30	2	2	2	3	3	3	5
6	3	3	3	7	3	3	5
9	3	3	3	10	3	3	5
16	3	3	3	11	3	3	5
16	3	3	3	19	3	3	5
18	3	3	3	22	3	3	5
28	3	3	3	8	3	4	5
12	2	4	4	14	3	4	5
31	2	4	4				

Figure 1 depicts mean point-to-point consumption of meters. While the meters that were clustered in clusters 2 and higher appear to have comparable consumption, the two meters 17 and 20 identified by clustering methods to belong to one cluster (1) appear to have consumption much higher than consumption of other meters, as well as a fairly unusual load shape.

Figure 1. Comparison of Consumption Plots for a Single Day



Clustering of Days in Data Set

In this section we present the results of clustering analysis as applied to daily load profiles of a collection of similar buildings. The intent is to automatically divide the data set into groups of similar daily consumption patterns. The analysts can then pay particular attention to a specific cluster that exhibits "anomalous" load profiles.

The data used for clustering analysis to identify unusual days were extracted from daily profiles of 31 meters. As mentioned above, the data sets covered the period from July 1999 through August 2001. Because some of the energy data was missing from the data set, we used only the days for which the entire 96 data points were available. Once again, the energy consumption was normalized by the square footage of the building.

We ran three experiments. The first experiment involved meter 17, which was identified in clustering buildings' consumption analysis as running with unusually high energy consumption. The second experiment involved a set of 6 meters, which were identified by clustering buildings' consumption analysis as running with low energy consumption (relative to energy consumed by other meters in the data set). The third experiment involved a meter (19) recommended by an analyst as a potentially interesting building. In all three experiments the data were clustered five ways.

For this analysis, we again used a K-means clustering algorithm with the Euclidean distance as the specified metric for calculating similarity.

Clustering results for meter 17. Clusters 2 and 4 obtained for meter 17 are shown in Figures 2 and 3 respectively. A number of plots are included in each figure. The upper plot represents the point-per-point median, mean, 25-percentile and 75-percentile consumption of the group of all meter-days in each cluster. The other 3 plots in a clock-wise direction represent histograms of days of the week, months of the year and years of the group of all meter-days in each cluster.

An interesting fact is that while the daily consumption of meters grouped in cluster 2 is similar to daily consumption of meters in clusters 1, 3, and 5, the daily consumption in cluster 4 is very different from daily consumption in other clusters. Cluster number 4 mostly represents summer-winter transition months for year 1999 on Tuesday through Sunday days of the week. In this cluster the off-hours consumption is higher than off-hours consumption in other clusters. In addition, in this cluster off-hours consumption is not very different from on-hours consumption. Since days represented by cluster number 4 are mostly summer-early fall days, the nightly energy consumption should differ from consumption during business hours. It appears that we identified a cluster that represents operation of the store attributable to a particular manager closing on Monday through Friday evenings.

Figure 2. Second Cluster for Meter 17

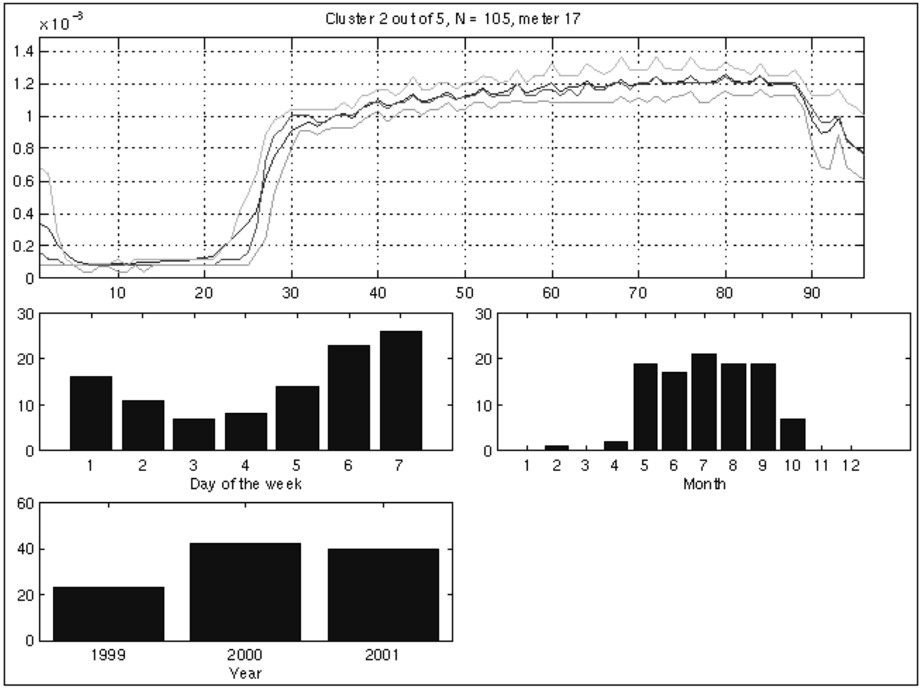
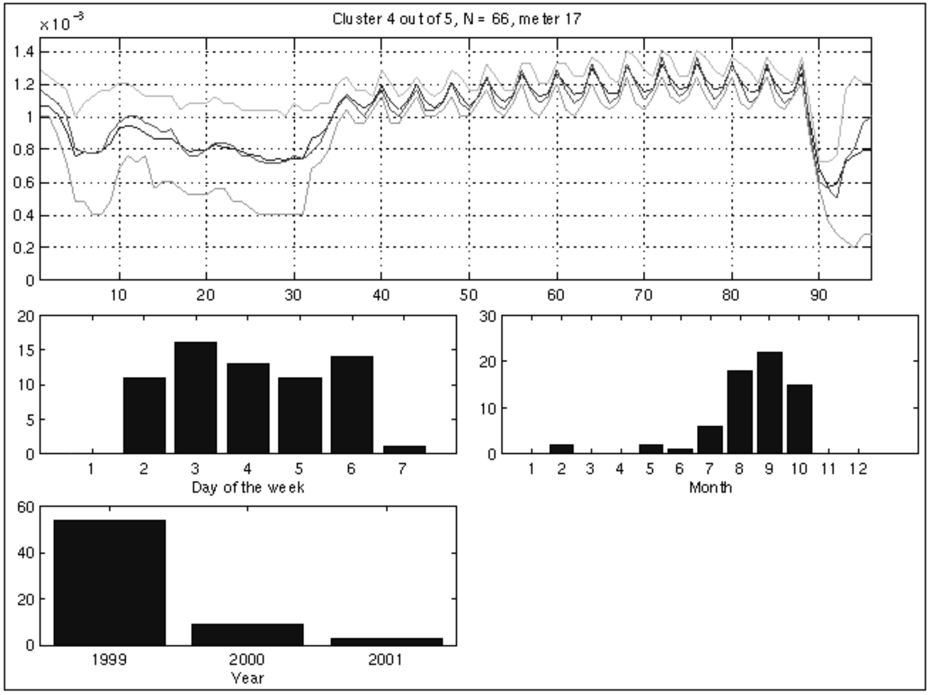


Figure 3. Fourth Cluster for Meter 17



Clustering results for group of six similar meters. For this experiment we used the following meters: 1, 13, 21, 23, 29, and 30. The resulting clusters 1, 4, and 5 are presented in figures 4 through 6. The daily consumption of meters in clusters 2 and 3 is similar to daily consumption of meters in cluster 1. The information in these figures is basically the same as information shown in figures 2 and 3 representing the clustering results of meter 17 with the addition on a plot in right lower corner of each figure that shows meter ID histogram in each cluster.

An interesting fact is the daily consumption in clusters 4 and 5 are different from the consumption in clusters 1, 2, and 3. The data in clusters 4 and 5 mostly represent the daily consumption measured by meter 13 during summer and winter months respectively. Cluster 4 differs from cluster 5 by having a greater difference between energy consumption during business hours and consumption during off-hours. This can be explained by the fact that cluster 4 represents summer months which require use of air conditioning and thus higher energy consumption. Overall, energy consumption in clusters 4 and 5 during off-hours is much higher than nightly consumption represented by data in clusters 1, 2, and 3.

Figure 4. First Cluster for Group of Six Similar Meters

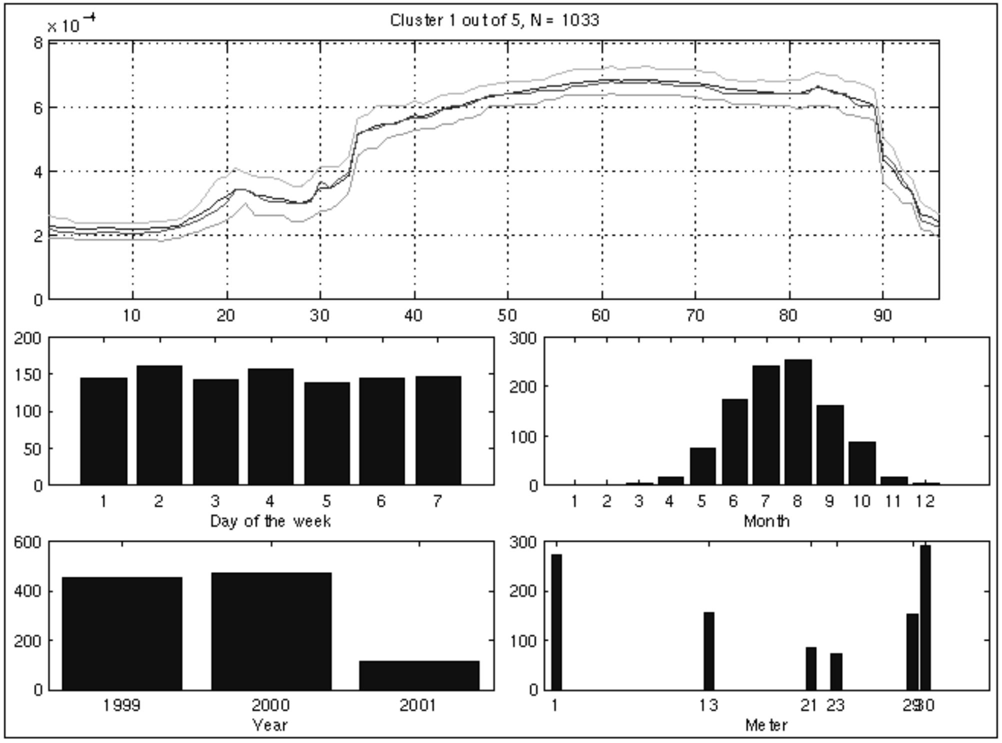


Figure 5. Fourth Cluster for Group of Six Similar Meters

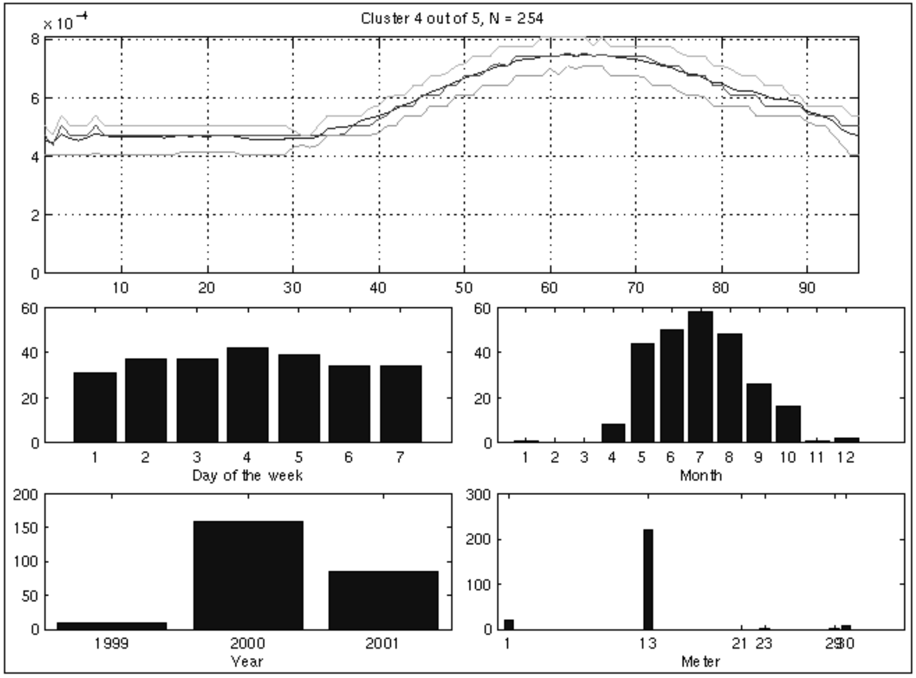
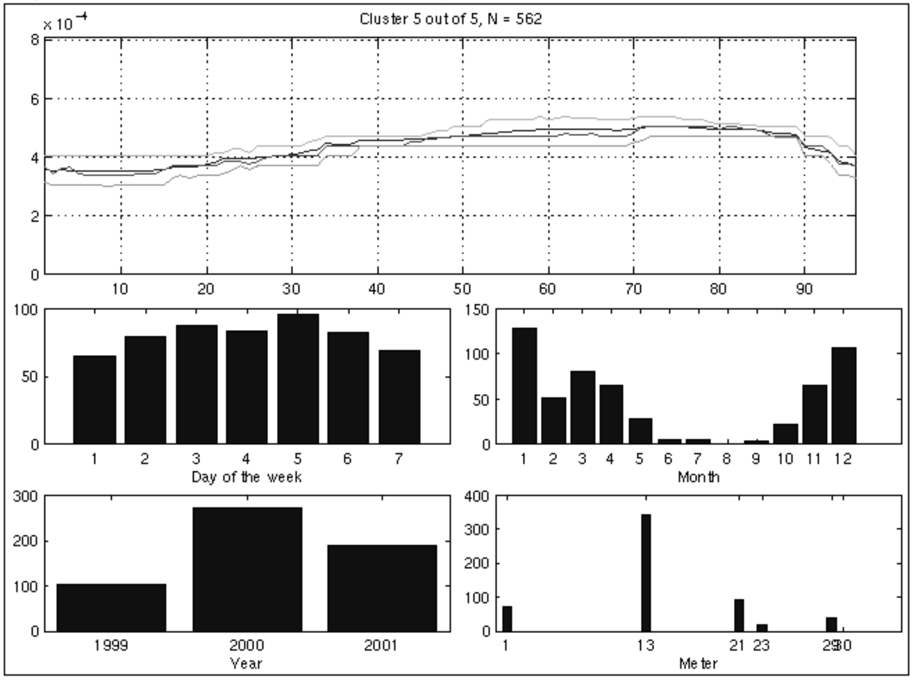


Figure 6. Fifth Cluster for Group of Six Similar Meters



Clustering results for meter 19. Clusters 1 and 2 are presented in Figures 7 and 8 respectively. The daily consumption of meters in clusters 3, 4, and 5 is similar to daily consumption of meters in cluster 1. The energy consumption in cluster 1 is different from the days represented by other clusters. First, the off-hours consumption in this cluster is higher than nightly consumption in other clusters. Second, the off-hours consumption is very close to business hours consumption in this cluster. Cluster 1 mostly represents summer-fall transition months of year 1999 on Tuesday through Sunday days of the week. Similarly to the results of cluster meter 17, it appears that clustering analysis was able to identify a cluster representing operation of the store by the particular manager closing on Monday through Friday evenings during 1999 summer-fall transition months.

Figure 7. First Cluster for Single Meter 19

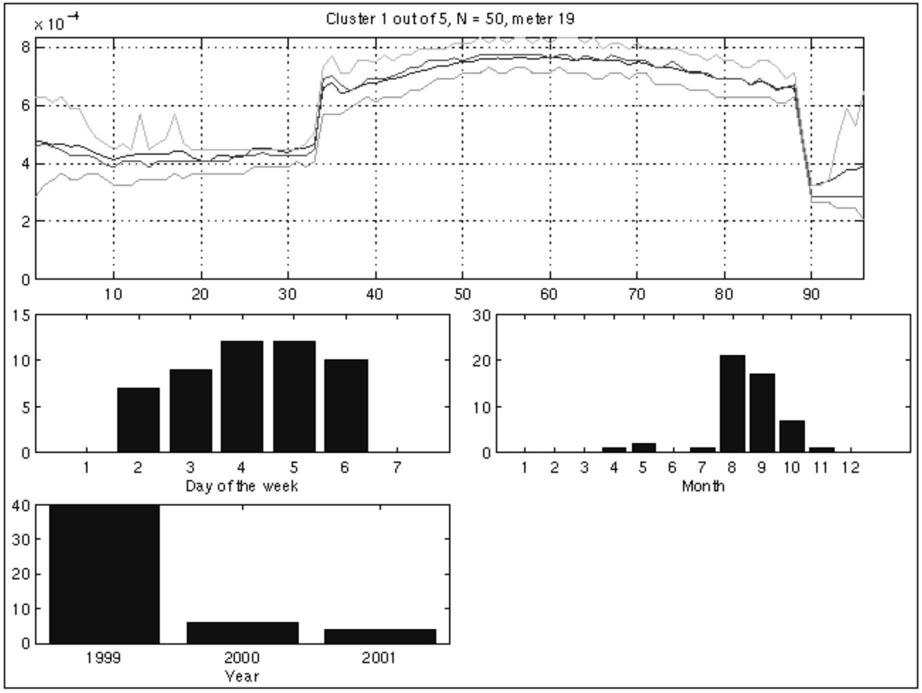
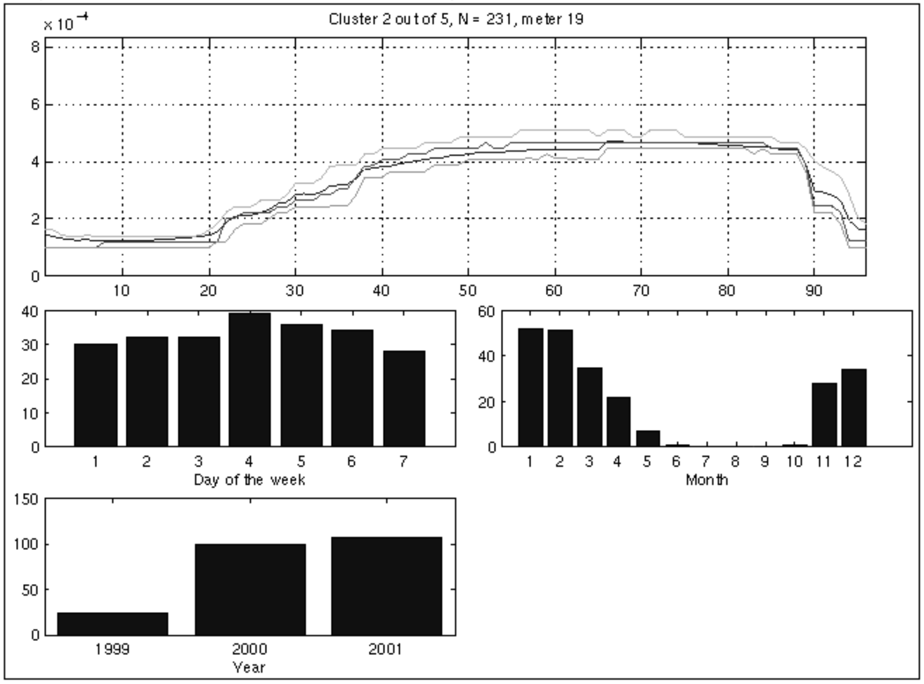


Figure 8. Second Cluster for Single Meter 19



Conclusions and Future Directions

In this paper, we have presented an approach to automating the analysis of large repositories of time series energy data. We have shown two approaches to clustering. The first approach used generalized energy profiles from a set of buildings as the population to be clustered. This approach demonstrated the ability to detect gross variations in building operation. These tend to be the type of variations that are relatively easy to detect by a human, given access to the data set. The second approach used actual daily energy profiles from a selected set of buildings as the population to be clustered. This approach gives insight into less obvious variations, such as changes in performance that are correlated to specific days of the week. The initial results from the clustering analysis are promising, but research issues need to be addressed before this approach is ready for widespread deployment.

Our experiments using data sets described below show that clustering analysis is a useful tool in guiding analysts to identify anomalous buildings and daily profiles and in enhancing analysts' productivity.

The preliminary results described in this paper appear to be promising. However, there is still significant work that needs to be done before trying to fully automate the clustering of Atrium™ data. We identified several research directions that we would like to investigate.

First, the data used to cluster facilities covered only 2 years. This amount of data did not allow a Fourier transform to identify features corresponding to retail cycles (for example Christmas gift buying). In the future we might encounter data sets that will be large enough for the Fourier transform to pick up features other than daily ones. The Fourier coefficients of different features will most likely differ in amplitude. We would like to be able to transform the feature space of Fourier coefficients in such a way that will allow accurate clustering.

Second, we would like to develop a more robust technique to cluster days that will be tolerant of noise. When clustering is applied to raw data, noise can potentially change the resulting clusters. In the future, we would like to be able to extract features from the raw daily profiles, similarly to the Fourier transform in clustering of individual meters, and use those features for clustering of consumption profiles. Feature extraction will not only remove noise from the data set, but can potentially reduce dimensionality of the data, which will make clustering more efficient. Efficiency of the clustering algorithm can become an issue as the size of the data set grows.

Third, we would like to remove temperature factors from the data. It is clear that some variation between clusters can be simply attributed to summer temperatures versus winter temperatures (for example cluster 4 and 5 in the group of six meters). By temperature de-trending the data sets we will be able to identify clusters of daily consumption which are driven by operational variables not directly related to weather, such changes in personnel.

Finally, we noticed that a problematic day could be characterized by having unusually high off-hours energy consumption. Unfortunately, off-hours consumption is usually lower than business hour consumption. Because the Euclidean distance is calculated on the raw data points, the distance between two days is usually driven by consumption during the day. There are a number of ways we can address this problem. First, we can weight the distance metric so that we emphasize the off hours behavior. Also, feature extraction and temperature de-trending may help to solve this problem.

References

R.O. Duda and P.E. Hart. 1973. *Pattern Classification and Scene Analysis*. John Wiley & Sons.

L. Kaufman and P.J. Rousseeuw. 1990. *Finding Groups in Data: an Introduction to Cluster Analysis.* John Wiley & Sons.

Richard Wesley Hamming, 1977, *Digital Filters*, Prentice-Hall Signal Processing Series, Englewood Cliffs, New Jersey

A.V. Oppenheim and R.W. Schafer, 1975, Digital Signal Processing, Prentice-Hall Signal Processing Series, Englewood Cliffs, New Jersey