An In-depth Study of Forecasting Household Electricity Demand using Realistic Datasets

Chien-Yu Kuo
Department of Computer Science and Information Engineering
National Taiwan Normal University
60047072s@ntnu.edu.tw

Ming-Feng Lee
Green Energy and Environment Research Lab
Industrial Technology Research Institute
mingfenglee@itri.org.tw

Chia-Lin Fu
Green Energy and Environment Research Lab
Industrial Technology Research Institute
fjlstar@itri.org.tw

Yao-Hua Ho
Department of Computer Science and Information Engineering
National Taiwan Normal University
yho@ntnu.edu.tw

Ling-Jyh Chen
Institute of Information Science
Academia Sinica
cclljj@iis.sinica.edu.tw

ABSTRACT

Data analysis and accurate forecasts of electricity demand are crucial to help both suppliers and consumers understand their detailed electricity footprints and improve their awareness about their impacts to the ecosystem. Several studies of the subject have been conducted in recent years, but they are either comprehension-oriented without practical merits; or they are forecast-oriented and do not consider per-consumer cases. To address this gap, in this paper, we conduct data analysis and evaluate the forecasting of household electricity demand using three realistic datasets of geospatial and lifestyle diversity. We investigate the correlations between household electricity demand and different external factors, and perform cluster analysis on the datasets using an exhaustive set of parameter settings. To evaluate the accuracy of electricity demand forecasts in different datasets, we use the support vector regression method. The results demonstrate that the medium mean absolute percentage error (MAPE) can be reduced to 15.6% for household electricity demand forecasts when proper configurations are used.

Categories and Subject Descriptors

H.4 [**Information Systems Applications**]: Miscellaneous

Keywords

Electricity demand forecast; household electricity demand; data analysis

e-Energy'14, June 11–13, 2014, Cambridge, UK.

ACM 978-1-4503-2819-7/14/06 ...$15.00.
http://dx.doi.org/10.1145/2602044.2602055.

1. INTRODUCTION

With recent advances in the Internet of Things (IoT) and smart grid technologies, smart electricity meters have been developed and are now widely deployed. Unlike conventional electricity meters that require labor-intensive reading, smart electricity meters exploit modern data communication techniques to transmit their readings to remote data centers periodically. The resulting automated meter reading (AMR) infrastructure provides more insightful information about electricity demand at a finer granularity. The information enables power suppliers to improve their electricity services, such as electricity billing, pricing, provisioning, and real-time demand responses. Moreover, it allows consumers to monitor their electricity consumption continuously, and it induces behavior changes that save energy and lead to more environmentally friendly lifestyles [2, 19, 25].

Several attempts have been made to analyze AMR data for value-added and advanced electricity services in recent years [6, 8, 10, 11, 12, 14, 15, 22, 25, 27]. These works can be categorized into two types based on their objectives, namely, *comprehension-oriented* studies and *forecast-oriented* studies. While the former focus on gaining a thorough understanding of the intrinsic properties of AMR data, the latter go one step further and forecast the demand for electricity. The drawback of existing works is that they only consider regional scenarios (e.g., commercial buildings, cities, and country areas) without considering the scenarios on a per-household basis. In fact, forecasting household electricity demand is regarded as challenging because it must take various human factors into account (e.g., income levels, activities, and lifestyles). A large-scale in-depth study of forecasting electricity demand in a household-based setting is therefore highly desirable.

In this study, we investigate the research problem of large-scale AMR data analysis with the objective of improving the accuracy of forecasting household electricity demand. Using the realistic datasets compiled by the Pilot Smart Meter

Deployment Project in Taiwan, as well as data on weather conditions recorded by the weather stations near the deployment areas, we analyze the AMR data and investigate its correlations with various external factors, such as the temperature, the number of the floor where the smart meter is located, and different time scales. Then, we use the support vector regression (SVR) method [26] to evaluate the accuracy of electricity demand forecasts under different parameter settings. The contribution of this work is three-fold.

1. We conduct in-depth analysis using realistic smart electricity meter datasets and investigate the correlations between electricity demand and several external factors, such as the time, temperature, and number of the floor where the meter is located. The datasets are large-scale (i.e., 1,296 meters), and cover a long period (i.e., more than two years). They are also diverse in terms of household types (i.e., apartments and houses) and locations (i.e., entirely residential areas and mixed residential/commercial areas).

2. We perform cluster analysis on the datasets under an extensive set of parameter settings (e.g., under different distance measurement methods, clustering algorithms, and cluster validity indices). In addition, we exploit a decision mechanism based on the Memetic Algorithm [20] to determine the optimal number of clusters for each configuration.

3. To evaluate the accuracy of forecasting household electricity demand, we apply the SVR method on different datasets under different parameter settings, and analyze the results in detail.

Based on our research findings, we draw the following conclusions.

1. There is a *turning temperature*, which means the daily household electricity demand and the daily average temperature have a positive correlation when the temperature is greater than the turning temperature; otherwise, they have a negative correlation.

2. There is a significant drop in the daily household electricity demand on non-working weekdays (holidays), but there are no significant variations in the demand on weekdays or weekends.

3. The datasets exhibit a *lifestyle diversity*, and the best forecast performance of each dataset occurs under different parameter settings. Overall, the medium MAPE of the best forecast achieved was 15.6% when proper configurations were used in this study.

The remainder of this paper is organized as follows. Section 2 contains a review of recent studies on smart meter data analysis. In Section 3, we describe the three datasets used in this study; and in Section 4, we analyze the datasets and their correlations with different external factors. In Section 5, we conduct cluster analysis on the datasets, and also describe the Memetic Algorithm-based approach used to determine the optimal number of clusters for each dataset under different parameter settings. In Section 6, we evaluate the accuracy of forecasting electricity demand forecast with the three datasets, and present a detailed analysis on the results. Section 7 contains some concluding remarks.

2. RELATED WORK

There have been a number of competitions to determine the best way to forecast demand. One of the most notable was the Electricity Load Competition hosted by EUropean Network Intelligent TEchnologies for Smart Adaptive Systems (EUNITE) in 2001 [1]. The contest utilized a dataset provided by the East-Slovakia Power Distribution Company with 578,082 take-off points in the Eastern Slovakian Region. The dataset contains the per-30-minute electricity loads and the daily average temperature between January 1997 and December 1998. Based on the dataset, the competitors were asked to predict the maximum daily electricity load for January 1999.

In the contest, several competitors observed that there was a strong correlation between the temperature and the electricity load. However, it has been found that the temperature feature is not useful in forecasting the electricity load unless the temperature forecast is accurate [8, 12, 14, 15]. Moreover, the electricity load is higher on weekends than on weekdays [10, 14]. The variation between weekdays and weekends can be exploited to improve the accuracy of electricity load forecasts [6, 8, 22]

Jain and Satish [11] conducted an electricity load forecast study using a 2-year dataset that contained the per-30-minute regional electricity load, the daily average temperature, and the day of week (DOW). They divided the dataset into several clusters based on arbitrary thresholds and used the SVR approach to make forecasts. Specifically, SVR predicted the electricity loads (i.e., 48 thirty-minute loads) based on the previous 48 thirty-minute loads, the previous day's average temperature, the DOW of the previous day, and the temperature forecast for that day.

Shen et al. [25] proposed a Pattern Forecasting Ensemble Model (PFEM) that combines the Pattern Sequence-based Forecasting (PSF) algorithm [17] with five clustering models (namely, K-Means, K-Medoids, Self-Organizing Map, Hierarchical Clustering, and Fuzzy C-Means) with different weights to derive more accurate electricity load predictions. The model was evaluated on three publicly available electricity demand datasets compiled by the New York Independent System Operator (NYISO), the Australian National Electricity Market (ANEM), and Ontario's Independent Electricity System Operator (IESO), respectively. The evaluation results showed that PFEM could provide more accurate and reliable forecasts than PSF with a single clustering method.

Finally, Solomom et al. [27] conducted a study to forecast the electricity demand of a large commercial building at 345 Park Avenue in Manhattan. Approximately 5,000 people work in the building and there are about 1,000 visitors every day. The authors used a dataset of the electricity demand in the building over fifteen months, The forecast, which was SVR-based, predicted the electricity demand for the next week based on previous electricity demands, daily average temperatures, dew point temperatures, and wind speed data.

3. DATASET

In this section, we present the datasets used in the current study and discuss their basic properties. The datasets were obtained from the Pilot Smart Meter Deployment Project in Taiwan, which was launched in 2010 with the support of the

Table 1: Summary of the dataset collected from the different deployment sites

Dataset (Site)	Type	Area	# of smart meters		Period (days)
			w/o floor info.	w/ floor info.	
Taipei	apartments	residential/commercial	715	540	992 (2010/11/08 - 2013/07/26)
Hsinchu	apartments	residential	419	240	999 (2010/11/01 - 2013/07/26)
Tainan	houses	residential	162	162	262 (2012/11/07 - 2013/07/26)

Taiwan Power Company and the Bureau of Energy, Ministry of Economic Affairs, Taiwan. The deployment started in the cities of Taipei and Hsinchu in the north of Taiwan; and in 2012, another deployment was added in Tainan, a city in the south of Taiwan.

The Taipei dataset comprises 715 smart electricity meters deployed in the Minsheng Community, which is a mixed-use area of residential and commercial properties in the east of Taipei City. Each smart meter corresponds to one household, and it reports the householdÕs electricity demand every fifteen minutes via different techniques (e.g., Wi-Fi and power line communication). Of the 750 smart meters listed in the dataset, 540 show the full postal address of the household (including the floor number) where they are installed. They are categorized into LOW floors (1F-5F with 464 meters); MIDDLE floors (6F-10F with 48 meters); and HIGH floors (11F-16F with 26 meters).

The Hsinchu dataset comprises 419 smart electricity meters deployed in the Siangshan Community in the southwest of Hsinchu County; again, each smart meter is associated with one household. The Siangshan Community is an entirely residential area, where most of the residents are either students or they work in nearby companies. The smart meters have similar configurations to those deployed in Taipei city including the sample rate and communication techniques. Of the 419 smart meters in the dataset, 240 show the complete postal addresses where they are located. There are 150 meters on LOW floors (1F-5F), 50 meters on MIDDLE floors (6F-10F), and 40 meters on HIGH floors (11F-18F), respectively.

The Tainan dataset only contains 162 smart meters, each of which is associated with one floor of a three-floor household (i.e., there are 54 households in the dataset). The complete postal addresses of the participating households are available. All the houses are located in the same community and have the same design: the living room is on the first floor (LOW floor), the kitchen and one bedroom are on the second floor (MIDDLE floor), and another two bedrooms are on the third floor (HIGH floor). Most of the residents are workers and students in nearby areas. The smart meters are configured to use the same sample rate and communication techniques as those in the Taipei and Hsinchu sites.

The deployment project was implemented in an incremental manner. In the early phase, the number of *concurrently alive meters* varied significantly over time due to meter failures and unreliable wireless communications. To resolve the problem, all the smart meters have been upgraded to use power line communications to transmit data, and the smart meters installed in the early phase of the project have been replaced with new models. Moreover, we designed a simple filter to remove obvious outliers from the datasets, such as negative electricity demands and extremely large demands (i.e., more than 100 kilowatt-hours) in a 15-minute time slot. Table 1 summarizes the datasets used in this study, and Figure 1 shows the number of the *concurrently alive meters* over time in the datasets.

To investigate the correlation between environmental factors and household electricity demand, we downloaded weather condition data from the weather station at Mingchuan elementary school, which is about one kilometer from the deployment site. For the Hsinchu and Tainan sites, we purchased weather condition data from the Taiwan Central Weather Bureau. We also obtained data from two weather observation stations that are about ten kilometers from the Hsinchu site and the Tainan site respectively.

4. DATA ANALYSIS

Next, we analyze the intrinsic properties of the three datasets; and then evaluate the correlations between electricity demand and various external factors.

Figure 2 shows the cumulative distribution function (CDF) of the data instances (i.e., *per-15-minute* electricity demand) in the three datasets. The curve of the Taipei dataset is always on the right of the other curves, indicating that the households in Taipei generally consume more energy than those in Hsinchu and Tainan. There are two reasons for this phenomenon: 1) some of the smart meters in the Taipei dataset are installed in premises used for commercial activities, which consume more power than strictly residential activities; and 2) the income level of households in Taipei is higher than that in the other two cities, so the residents tend to consume more energy [4]. In addition, we observe that 58% of the data instances in the Hsinchu dataset have extremely low values (i.e., less than 0.05kWh). This is because most of the households in the Hsinchu site are either single-person or dual-income-no-kids (DINK) entities. The residents tend to work long hours, even at weekends, in high-tech companies in the nearby Hsinchu Science Park; hence, the demand for electricity is extremely low.

Figure 3 compares the per-15-minute electricity demand of LOW, MIDDLE, and HIGH floor households over a day in the three datasets. We make the following observations.

1. Overall, electricity demand is highest at night and lowest in the early morning. The reason is simple: the demand is highly correlated to people's household activities at different times of the day.

2. In the Taipei dataset, LOW-floor households consume more power than the other groups around noon (11am - 1pm); while HIGH-floor households consume more power than the other groups in the evening (8pm - 9am). The former can be explained by the fact that some LOW-floor premises are used for business, and they are busier during the lunch period than at other times. The reason for the latter finding is that the temperatures in Taipei households on HIGH floors usually differ substantially from those on LOW floors due to the basin effect in Taipei. Thus, people on HIGH floors

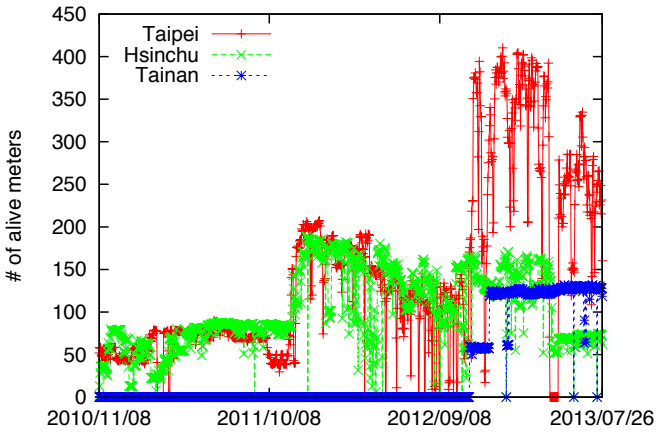


Figure 1: The number of the *concurrently alive meters* over time in the three datasets

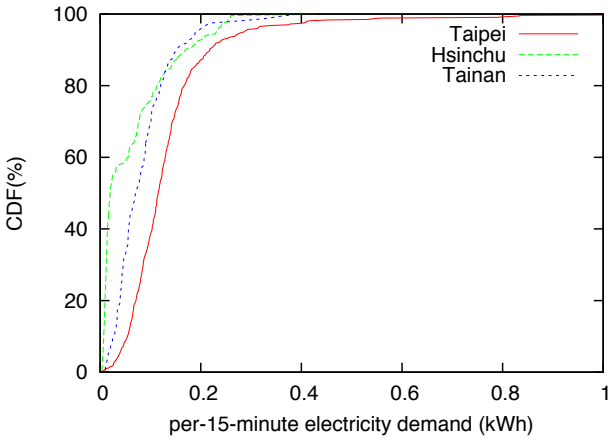
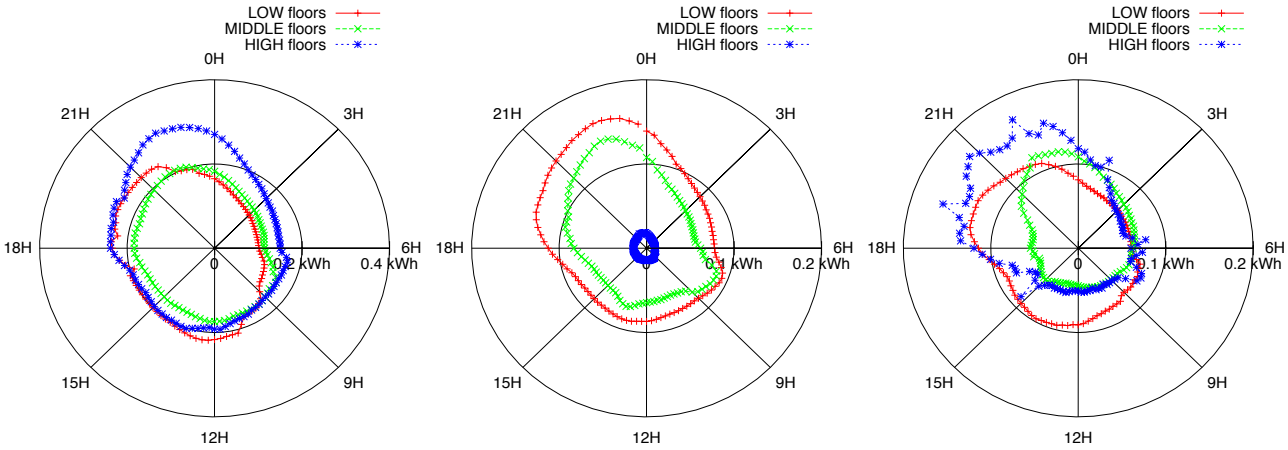


Figure 2: The CDF of the *per-15-minute* electricity demand in the three datasets



(a) Taipei (b) Hsinchu (c) Tainan

Figure 3: Comparison of the per-15-minute electricity demand on the LOW/MIDDLE/High floors in a day at the three sites

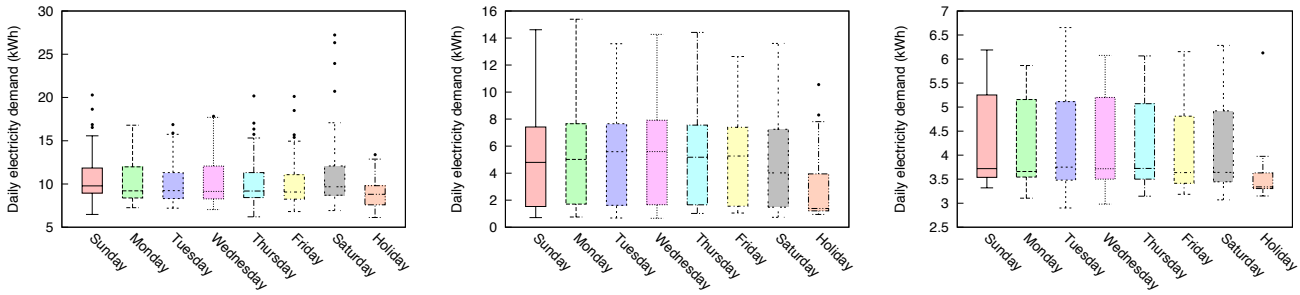
tend to use air conditioners more often in summer, and turn on heaters and dehumidifiers more often in winter.

3. In the Hsinchu dataset, LOW-floor households consume more energy than MIDDLE-floor premises over a day; while HIGH-floor households consume the least amount. The reason is that the deployment is located in a new residential area and many premises are not currently occupied. Interestingly, higher floors have more vacancies than middle and lower floors.

4. In the Tainan dataset, first-floor households consume the most electricity during the day, while third-floor households have the highest demand at night. The finding shows that people usually spend most of their time in the living room (1F) during the day, while the bedrooms (2F and 3F) are used mostly at night.

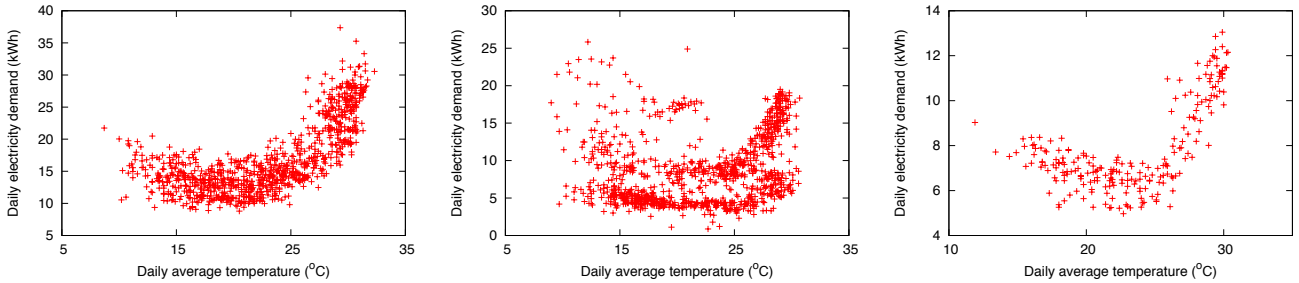
We also evaluate the correlations between the per-15-minute electricity demand and different types of day, i.e., working weekdays, weekends, and non-working weekdays (i.e., holidays). The results in Figure 4 show that electricity demand on non-working weekdays is lower than on the other days. This is because most people go out (or go away) on those days. Moreover, the electricity demand on working weekdays is comparable to that on weekends in the three datasets. The reason is that people may also go out on weekends (e.g., for leisure instead of work), so their household life patterns are the same on working weekdays and weekends.

In addition, we investigate the impact of the daily average temperature on the daily electricity demand registered by each smart meter. From the results shown in Figure 5, we observe that the distribution forms a *V shape* in all three datasets. More precisely, there is a *turning temperature* in each dataset (i.e., the valley of the *V shape*); the greater the difference between the daily average temperature and the turning temperature, the higher the electricity consumption. The reason is that people tend to turn on air conditioners in summer and heaters in winter. Thus, the turning temperature is deemed to be the ideal temperature for most people (i.e., they feel comfortable) so there is no need for air conditioners/heaters. Specifically, the results in Figure 5 show



(a) Taipei (b) Hsinchu (c) Tainan

Figure 4: Comparison of the per-15-minute electricity demand on working weekdays, weekends, and non-working weekdays (i.e., holidays) in the three datasets



(a) Taipei (b) Hsinchu (c) Tainan

Figure 5: Illustration of the distribution of the daily electricity demand under different daily average temperatures in the three datasets

that the turning temperatures are 20^oC, 25^oC, and 25^oC in the Taipei, Hsingchu, and Tainan datasets respectively.

Finally, using the turning temperature in Figure 5, we evaluate the correlations between the daily average temperature and the daily electricity demand for households on different floors (i.e., HIGH/MIDDLE/LOW floors) under different weather conditions (i.e., above/below the turning temperature). The results in Table 2 show that, when the temperature is higher than the turning temperature, the correlation is strongly positive in the Taipei and Tainan datasets, and moderately positive in the Hsinchu dataset. In contrast, when the temperature is lower than the turning temperature, the correlation is weakly negative except for those household on MIDDLE and HIGH floors in the Taipei and Tainan datasets, which have moderately negative correlations. The reason is that heaters are not always needed during winter in the three cities because the climate in Taiwan ranges from subtropical in the north to tropical in the south. Thus, the correlation is moderate to weak when the daily average temperature is lower than the turning temperature.

Table 2: Correlations between the daily average temperature and the daily electricity demand for smart meters on different floors and under different weather conditions in the three datasets

		LOW floors	MIDDLE floors	HIGH floors
Taipei	$\geq 20^oC$	0.780	0.823	0.783
	$< 20^oC$	-0.227	-0.654	-0.392
Hsinchu	$\geq 25^oC$	0.572	0.331	0.590
	$< 25^oC$	-0.288	-0.198	-0.094
Tainan	$\geq 25^oC$	0.831	0.869	0.757
	$< 25^oC$	0.194	-0.583	-0.655

5. CLUSTER ANALYSIS

We use cluster analysis to identify households with similar lifestyle patterns in the three datasets. In the analysis, we use the *daily* electricity demand, instead of per-15-minute consumption, because it is more representative of the seasonal changes in people's household lifestyle patterns. Moreover, we evaluate different parameter settings (e.g., distance measures, clustering algorithms, and cluster metrics) and implement a Memetic Algorithm-based approach to determine the optimal number of clusters under each parameter setting. We present the detailed analysis in the following subsections.

5.1 Data Preprocessing

Two issues must be resolved before performing cluster analysis on the three datasets: 1) *data dependency*: the daily electricity demand is the combined result of people's lifestyle patterns and external factors, such as the temperature (as shown in Table 2); and 2) *data scaling*: different households may have similar behavior patterns in terms of daily electricity demand, but on different scales. The demand depends on the number (and the models) of the appliances used in different households. Thus, it is necessary to preprocess the datasets in order to mitigate the effects of the two issues. The data preprocessing phase involves two steps:

- *Data Separation*
  Using the turning temperatures discussed in Section 4,

we divide each dataset into two subsets: (1) the *warm subset*, which contains data instances for the months in which the average temperatures were higher than the turning temperature; and (2) the *cold subset*, which contains the rest of the data instances in the original dataset. In this study, the warm seasons are defined as April to November for the Taipei dataset, May to October for the Hsinchu dataset, and April to October for the Tainan dataset.

- *Data Standardization*
To resolve the data scaling issue, we exploit the classic data *standardization* approach to convert the daily electricity demand from its raw form into a *standard score (z-Score)* [3]. Specifically, we let $X_k^{i,j}$ be the electricity demand recorded by the i-th smart meter in the k-th time slot of the j-th day in the dataset. The standard score of $X_k^{i,j}$ is

$$Z_k^{i,j} = \frac{X_k^{i,j} - \mu(X_*^{i,j})}{\sigma(X_*^{i,j})}, \tag{1}$$

where $\mu(X_*^{i,j})$ is the mean electricity demand of the i-th household on the j-th day, and $\sigma(X_*^{i,j})$ is the standard deviation of the electricity demand recorded by the i-th smart meter for all time slots on the j-th day.

5.2 Parameter Settings

In this subsection, we consider the different parameter settings used in the cluster analysis, including distance measures, clustering algorithms, and cluster validity indices. We discuss the possible settings of each parameter and the rationale for each one in the following subsections.

5.2.1 Distance Measure

Let $Z_k^{i,j}$ be the standardized electricity demand recorded by the i-th smart meter in the k-th time slot of the j-th day, and let n be the number of time slots in a day. We consider two distance functions to measure the distance between $Z_*^{i_1,j_1}$ and $Z_*^{i_2,j_2}$.

- *Euclidean distance:* The Euclidean distance represents the ordinary distance between two points in geometry, and it is widely used to measure the distance between two samples in a multi-dimensional space. In this study, we derive the Euclidean distance between $Z_*^{i_1,j_1}$ and $Z_*^{i_2,j_2}$ by computing the square root of the sum of the squares of the differences between two standardized electricity demands, i.e.,

$$D^{Euclid}(Z_*^{i_1,j_1}, Z_*^{i_2,j_2}) = \sqrt{\sum_{k=1}^{n} (Z_k^{i_1,j_1} - Z_k^{i_2,j_2})^2}. \tag{2}$$

- *Dynamic Time Warping (DTW) Distance:* The DTW distance [13] is commonly used to measure the similarity between two time sequence data events that may vary in time and speed. To calculate the DTW distance between $Z_k^{i_1,j_1}$ and $Z_k^{i_2,j_2}$, we first identify four time slots (as shown in Figure 6):

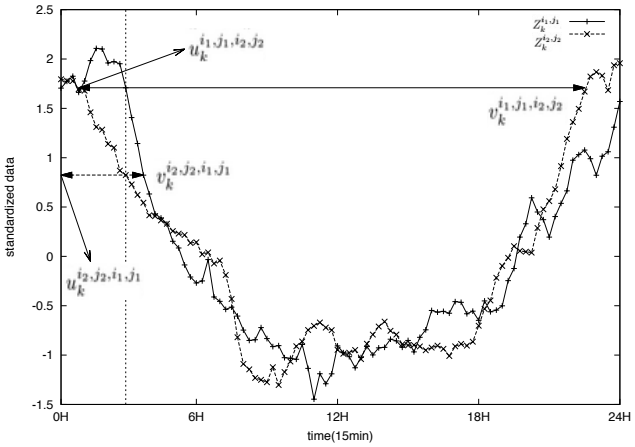


Figure 6: Illustration of the Dynamic Time Warping (DTW) distance

1. $u_k^{i_1,j_1,i_2,j_2}$: the nearest backward slot of $Z_*^{i_1,j_1}$ such that its value is equal to $Z_k^{i_2,j_2}$. i.e.,

$$u_k^{i_1,j_1,i_2,j_2} = \max\left(0, \underset{1\le k'<k}{\text{argmax}} (Z_{k'}^{i_1,j_1} = Z_k^{i_2,j_2})\right). \tag{3}$$

2. $v_k^{i_1,j_1,i_2,j_2}$: the nearest forward slot of $Z_*^{i_1,j_1}$ such that its value is equal to $Z_k^{i_2,j_2}$, i.e.,

$$v_k^{i_1,j_1,i_2,j_2} = \min\left(0, \underset{k<k'\le n}{\text{argmin}} (Z_{k'}^{i_1,j_1} = Z_k^{i_2,j_2})\right). \tag{4}$$

3. $u_k^{i_2,j_2,i_1,j_1}$: the nearest backward slot of $Z_*^{i_2,j_2}$ such that its value is equal to $Z_k^{i_1,j_1}$, i.e.,

$$u_k^{i_2,j_2,i_1,j_1} = \max\left(0, \underset{1\le k'<k}{\text{argmax}} (Z_{k'}^{i_2,j_2} = Z_k^{i_1,j_1})\right). \tag{5}$$

4. $v_k^{i_2,j_2,i_1,j_1}$: the nearest forward slot of $Z_*^{i_2,j_2}$ such that its value is equal to $Z_k^{i_1,j_1}$, i.e.,

$$v_k^{i_2,j_2,i_1,j_1} = \min\left(0, \underset{k<k'\le n}{\text{argmin}} (Z_{k'}^{i_2,j_2} = Z_k^{i_1,j_1})\right). \tag{6}$$

Then, the directional DTW distance between $Z_k^{i_1,j_1}$ and $Z_k^{i_2,j_2}$ (i.e., $\overrightarrow{D}^{DTW}(Z_k^{i_1,j_1}, Z_k^{i_2,j_2})$) and the distance in the opposite direction (i.e., $\overrightarrow{D}^{DTW}(Z_k^{i_2,j_2}, Z_k^{i_1,j_1})$) are obtained by

$$\begin{aligned} &\overrightarrow{D}^{DTW}(Z_k^{i_1,j_1}, Z_k^{i_2,j_2}) \\ = \;& \min\left(k - u_k^{i_1,j_1,i_2,j_2}, v_k^{i_1,j_1,i_2,j_2} - k\right), \text{ and} \end{aligned} \tag{7}$$

$$\begin{aligned} &\overrightarrow{D}^{DTW}(Z_k^{i_2,j_2}, Z_k^{i_1,j_1}) \\ = \;& \min\left(k - u_k^{i_2,j_2,i_1,j_1}, v_k^{i_2,j_2,i_1,j_1} - k\right). \end{aligned} \tag{8}$$

Finally, the DTW distance between $Z_*^{i_1,j_1}$ and $Z_*^{i_2,j_2}$ is determined by considering either the mean directional distance of each corresponding pair (as shown

in Eq. 9) or the minimum directional distance of each corresponding pair (as shown in Eq. 10).

$$D_{mean}^{DTW}(Z_k^{i_1,j_1}, Z_k^{i_2,j_2}) = \sum_{k=1}^{n} \frac{\overrightarrow{D}^{DTW}(Z_k^{i_1,j_1}, Z_k^{i_2,j_2}) + \overrightarrow{D}^{DTW}(Z_k^{i_2,j_2}, Z_k^{i_1,j_1})}{2}. \quad (9)$$

$$D_{min}^{DTW}(Z_k^{i_1,j_1}, Z_k^{i_2,j_2}) = \sum_{k=1}^{n} \min(\overrightarrow{D}^{DTW}(Z_k^{i_1,j_1}, Z_k^{i_2,j_2}), \overrightarrow{D}^{DTW}(Z_k^{i_2,j_2}, Z_k^{i_1,j_1})). \quad (10)$$

5.2.2 Clustering Algorithms

We exploit two clustering algorithms in the analysis: the K-Means algorithm [16] and the K-Medoids algorithm [24]. The K-Means algorithm is one of the most popular methods used in data mining. It partitions the data space into several Voronoi cells such that the distance between a data point and the geometric center of its own Voronoi cell is less than the distance to the centers of any other cells.

Specifically, given K as a priori information, the K-Means algorithm clusters the daily electricity demand dataset as follows:

Step 1: It selects K data instances at random from the dataset as the cluster centers.

Step 2: For the rest of the data instances in the dataset, it uses Eq. 2 to calculate the Euclidean distance from each data instance to each cluster center. Then, it associates each data instance with the closest cluster center.

Step 3: It calculates the geometric center of each cluster, and updates the cluster centers accordingly.

Step 4: It repeats Step 2 and Step 3 until no further changes can be made.

The K-Medoids algorithm is implemented in a similar way to the K-Means algorithm, except that: 1) in Step 2, it utilizes the DTW distance (Eq. 9 and Eq. 10) instead of the Euclidean distance to measure the distance between two data instances in the dataset; and 2) in Step 3, it chooses one of the samples as the new cluster center, such that the sum of the distances between the new cluster center and the other data instances in the same cluster is the minimum. Depending on the distance function used, the K-Medoids algorithm has two variants: K-Medoids-mean (i.e., using Eq. 9) and K-Medoids-min (i.e., using Eq. 10).

5.2.3 Cluster Validity Index

Instead of using an arbitrary number K as a priori number of clusters, we exploit two cluster validity indices, the PBM index [21] and the Davies-Bouldin (DB) index [9], to determine the optimal number of clusters for data clustering. Let K be the number of clusters; C_i be the i-th cluster; c_i be the center of the i-th cluster; and $z_{i,j}$ be the j-th data instance in the i-th cluster. We calculate the two cluster validity indices as follows.

- *PBM Index:* The *intra-cluster* distance between data instances belonging to the same cluster and the *inter-cluster* distance between two cluster centers are derived by Eq. 11 and Eq. 12 respectively; where $|C_i|$ denotes the number of data instances belonging to the i-th cluster, and D^* is the distance function used.

$$E_K = \sum_{i=1}^{K} \sum_{j=1}^{|C_i|} D^*(z_{i,j}, c_i). \quad (11)$$

$$F_K = \max_{1 \le i \ne j \le K} D^*(c_i, c_j). \quad (12)$$

Then, given a cluster number K, the PBM index is obtained by

$$PBM(K) = \left(\frac{1}{K} \times \frac{E_1}{E_K} \times F_K\right)^2, \quad (13)$$

where E_1 is the sum of the distances of all data instances to the center of the dataset. The larger the value of $PBM(K)$, the more stable will be the clustering results. The optimal number of clusters K is obtained when the value of $PBM(K)$ is maximal.

- *DB Index:* We derive the intra-cluster distance for the i-th cluster by Eq. 14. Then, we calculate the *goodness-of-fit* of the i-th cluster by finding the best partner cluster that can maximize the ratio of the intra-cluster distance over the inter-cluster distance, as shown in Eq. 15.

$$S_i = \frac{1}{|C_i|} \sum_{j=1}^{|C_i|} D^*(z_{i,j}, c_i). \quad (14)$$

$$R_i = \max_{1 \le j \le K; j \ne i} \frac{S_i + S_j}{D^*(c_i, c_j)}. \quad (15)$$

The DB index is obtained by calculating mean goodness-of-fit value of the K clusters, as shown in Eq. 16. In contrast to the PBM index, the optimal number of clusters K is obtained when the value of $DB(K)$ is minimal.

$$DB(K) = \frac{1}{K} \sum_{i=1}^{K} R_i. \quad (16)$$

5.3 Clustering Results

It has been shown that the cluster initialization issue influences the clustering results of many clustering algorithms [5, 28]. We address the issue as follows. Given a fixed number of clusters K, we utilize the Memetic-based algorithm [20] to combine *local search* and *genetic algorithms* to find the optimal clustering results. Figure 7 shows the flowchart of the algorithm.

There are eight steps:

- *Initial Population*: The algorithm selects K data instances at random as the initial cluster centers. Then, it repeats the process N_1 times to form N_1 sets of initial cluster centers.

- *Local Search*: For each set of cluster centers, the algorithm uses the selected clustering algorithm to analyze the dataset and yields K clusters.

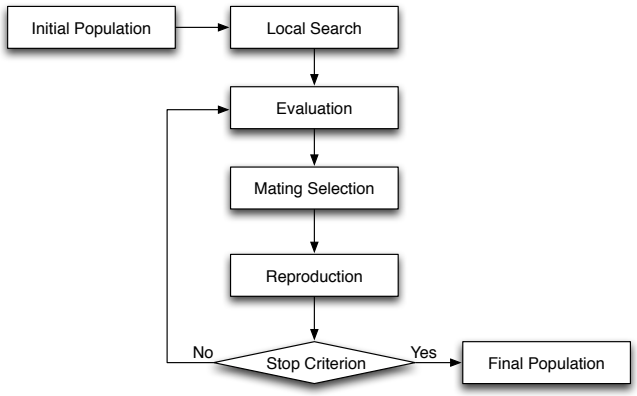


Figure 7: The flowchart of the Memetic-based algorithm used for cluster initialization

- *Evaluation*: The algorithm uses the selected cluster validity index to calculate the index value of the clustering results of each set of cluster centers.
- *Mating Selection*: The algorithm uses the *Tournament Selection* method [18] to select N_2 sets of cluster centers at random ($N_2 < N_1$), and reports the top two sets of clustering results (i.e., those that yield the maximal PBM index value or the minimal DB index value).
- *Reproduction*: Using the two sets of cluster centers derived by the *Mating Selection* step, the algorithm creates two sets of offspring by *one-point order crossover* [23] without mutation.
- *Environmental Selection*: The algorithm selects the best N_1 sets of cluster centers (from the initial N_1 sets and the two sets of offspring) as the new initial population for the next generation.
- *Stop Criterion*: The algorithm stops when it has iterated T times.
- *Final Population*: The algorithm outputs the best clustering results from the latest population.

Using the Memetic Algorithm with $N_1 = 10$, $N_2 = 4$, and $T = 20$, we performed cluster analysis on the three datasets under different parameter settings with different K values. Then, we obtain the optimal K value that yields the optimal cluster validity index value under each parameter setting for all the clusters. Table 3 shows the optimal K values under different clustering algorithms, distance measures, cluster validity indices and weather conditions in the three datasets.

6. FORECAST PERFORMANCE

Next, based on our analysis, we evaluate the accuracy of forecasts of the daily electricity demand for each smart meter. We use the ε-support vector regression (ε-SVR) method [26, 29] provided by the open source machine learning library, LIBSVM [7], as the forecasting tool. Specifically, in ε-SVR, the width of the ε-insensitive tube is set at 0.14286 (the default value suggested by the LIBSVM library); and the cost of errors is set at 4,096 for the Hsinchu dataset, 32 for the Taipei dataset, and 32 for the Tainan dataset.

Table 3: The optimal number of clusters suggested for each dataset under different clustering algorithms, distance measures, cluster validity indices and seasons

		PBM Index		DB Index	
		warm season	cold season	warm season	cold season
Taipei	K-Means	3	3	3	3
	K-Medoids-mean	2	2	4	3
	K-Medoids-min	5	5	3	4
Hsinchu	K-Means	3	3	3	3
	K-Medoids-mean	2	4	2	8
	K-Medoids-min	4	3	3	8
Tainan	K-Means	2	3	2	4
	K-Medoids-mean	3	2	5	3
	K-Medoids-min	4	5	2	5

Table 4: The number of smart meters suitable for forecast evaluations, when $d = 7$, in the three datasets

		Taipei	Hsinchu	Tainan
warm season	LOW floors	412	119	51
	MIDDLE floors	48	41	33
	HIGH floors	26	31	33
	Total	486	191	117
cold season	LOW floors	423	133	52
	MIDDLE floors	44	43	28
	HIGH floors	19	34	29
	Total	486	210	109

In addition, we perform *three-fold cross validation* for each dataset under different clustering parameter settings; and we use the mean absolute percentage error (MAPE) as the evaluation metric to measure the forecast accuracy for each smart meter. The MAPE value of the i-th smart meter is derived by

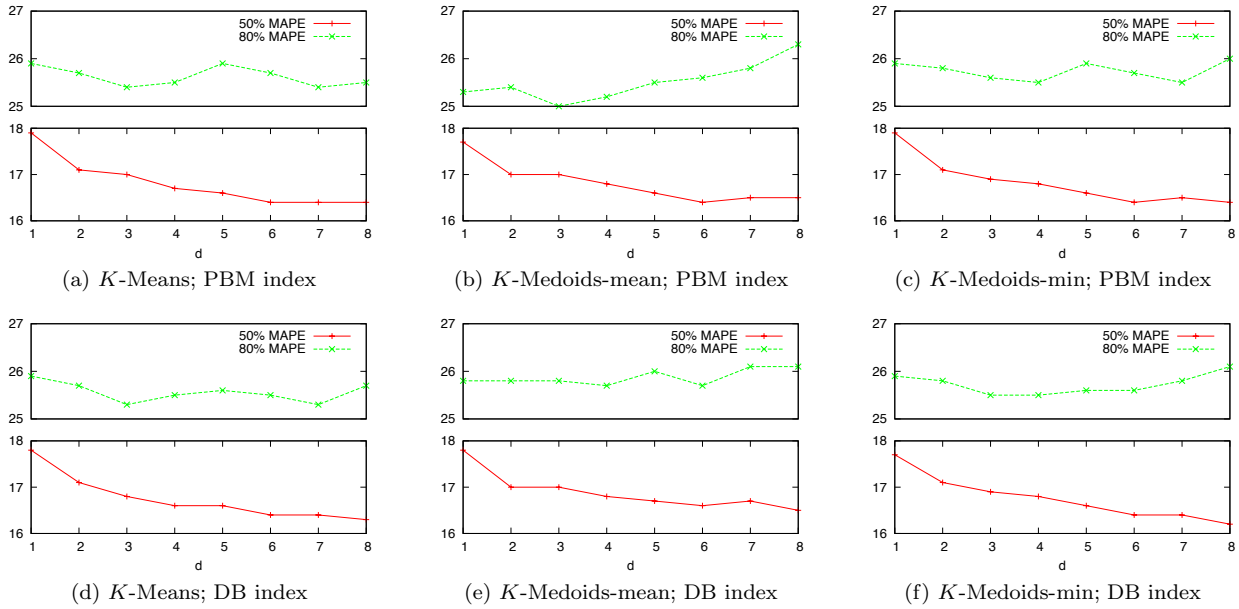
$$MAPE_i = 100 \times \frac{1}{p} \times \sum_{j=1}^{p} \frac{\left| V_{i,j} - \widehat{V_{i,j}} \right|}{V_{i,j}}; \qquad (17)$$

where $V_{i,j}$ is the true result of the j-th forecast on the i-th smart meter, $\widehat{V_{i,j}}$ is the j-th forecast value on the i-th smart meter, and p is the number of forecasts made for the i-th smart meter. The distribution of the MAPE values is skewed due to extreme values in general. Therefore, in the following discussion, we consider the 50% (i.e., median) and 80% MAPE values of all smart meters under different parameter settings.

6.1 Feature Selection

We use the *immediate last d contiguous days* of data instances (including the electricity demand and the other external factors) as the selected features for SVR forecasting. Intuitively, there exists a tradeoff in deciding the d value. The higher the value of d, the greater will be similarity between the lifestyles of the instances matched and the one to be forecast. However, using a large d value may reduce the number of matched instances, resulting in a failed or biased forecast.

In Figure 8, we compare the accuracy of forecasts of the daily electricity demand for households in the Taipei dataset with different d values, clustering algorithms, and cluster



(a) K-Means; PBM index (b) K-Medoids-mean; PBM index (c) K-Medoids-min; PBM index

(d) K-Means; DB index (e) K-Medoids-mean; DB index (f) K-Medoids-min; DB index

Figure 8: Comparison of the 50% and 80% MAPE results of forecasting daily household-based electricity demand using the Taipei dataset with different d values, clustering algorithms, and cluster validity indices

validity indices. We observe that when $d = 7$, the medium (50%) MAPE value is either the first or the second minimal value among the d values in all test cases. The 80% MAPE value is also approximately minimal when $d = 7$. The observations also apply to the Hsinchu and Tainan datasets, but we do not provide the MAPE values here due to the page limitation. Therefore, we set $d = 7$ in the following evaluation. Using the d value, Table 4 shows the number of smart meters in each dataset that have eight contiguous days of electricity demand and are eligible for the following evaluation.

6.2 Forecast Results

In addition to the daily electricity demand, we consider the following three external factors in the evaluation of the forecast performance: the temperature, the floor number, and the type of day. When the temperature feature is considered, we append the daily average temperature record to each daily electricity demand instance in the dataset. We also include this feature in the SVR training set and test set. When the floor number feature is considered, we divide the dataset into three subsets based on the floor type (i.e., HIGH, MIDDLE, or LOW) of each smart meter in the dataset. Moreover, when the type of day feature is considered, we divide the dataset into two subsets based on whether the measurement was obtained on regular days (working weekdays and weekends) or holidays (non-working weekdays).

Table 5 shows the 50% and 80% MAPE results of forecasting daily electricity demand using the three datasets with different cluster parameter settings and different feature combinations. Each feature is labeled either T (true) or F (false) based on whether it is considered, and the labels of the three features are concatenated in the following order: temperature, floor number, and day type. For instance, 'TFF' refers to a case where we only consider the temperature, and 'FTF' refers to a case where we only consider the floor number feature.

The results in Table 5 show that, for the Taipei and Hsinchu scenarios, the accuracy of the forecasts derived by the PBM and DB indices is comparable; the best forecast performance is achieved when the K-Means clustering algorithm and the Euclidean distance are used together. However, for the Taipei scenario, the best forecast performance only occurs when the type of day feature is considered (i.e., FFT). This result seems to contradict our earlier findings (discussed in Section 4) that the demand for electricity is highly correlated to the temperature. There are three reasons for this phenomenon:

1. Although the earlier analysis showed a strong correlation between electricity demand and the temperature feature, the distribution per temperature value spans a wide range of electricity demands (c.f., Figure 5), so forecasting demand is a difficult task.

2. The earlier analysis was performed by dividing the dataset into the warm season subset and the cold season subset. This procedure cannot be used to forecast electricity demand because the daily average temperatures in a contiguous d-day period may not belong to the same season subset.

3. Some of the smart meters in the Taipei dataset are located in premises used for commercial activities, which are more responsive to holidays. By contrast, the other two datasets only contain residential properties.

As a consequence, the temperature feature does not improve the accuracy of forecasts of the daily electricity demand in the Taipei scenario.

The results in Table 5 also show that the best forecast performance for the Hsinchu scenario is only achieved when

Table 5: The 50% and 80% MAPE results of forecasting daily electricity demand using the three datasets with different cluster parameter settings and different feature combinations

			FFF	FFT	FTF	FTT	TFF	TFT	TTF	TTT
Taipei	PBM Index	K-Means	16.4/25.4	***16.4/25.2***	16.6/26.3	16.9/26.1	16.6/26.2	17.1/26.4	17.2/27.3	17.7/27.9
		K-Medoids-mean	16.5/25.8	16.5/26.0	16.8/26.5	16.8/26.8	16.9/26.3	17.2/26.4	17.1/27.2	17.7/27.4
		K-Medoids-min	16.5/25.5	16.6/25.8	16.7/26.4	17.0/27.1	16.8/26.7	17.4/27.1	17.5/27.3	18.2/28.8
	DB Index	K-Means	16.4/25.3	***16.4/25.2***	16.5/25.9	16.8/25.9	16.6/25.9	17.0/25.9	17.1/26.9	17.7/27.7
		K-Medoids-mean	16.7/26.1	16.6/26.2	16.9/26.5	16.9/26.9	16.7/26.6	17.2/27.3	17.2/27.4	17.7/28.3
		K-Medoids-min	16.4/25.8	16.4/25.8	16.8/26.2	16.8/26.6	16.5/26.0	17.0/26.5	17.1/26.9	17.5/27.8
Hsinchu	PBM Index	K-Means	21.9/39.3	22.4/38.3	***21.9/37.7***	23.0/40.0	33.3/64.0	29.3/54.3	30.4/60.1	34.1/75.1
		K-Medoids-mean	22.0/38.9	22.4/39.3	22.3/39.1	22.8/38.0	32.0/63.7	28.1/46.5	33.6/68.1	33.8/59.3
		K-Medoids-min	22.2/39.0	22.7/38.6	22.4/38.0	23.2/38.8	32.2/62.4	30.6/60.1	32.6/60.0	34.6/75.4
	DB Index	K-Means	21.9/39.3	22.4/38.3	***21.9/37.7***	23.0/40.0	33.3/64.0	29.3/54.3	30.4/60.1	34.1/75.1
		K-Medoids-mean	22.0/38.9	22.4/39.3	22.3/39.1	22.8/38.0	32.0/63.7	28.1/46.5	33.6/68.1	33.8/59.3
		K-Medoids-min	21.9/37.8	22.2/38.4	22.1/38.4	23.2/37.0	33.5/59.0	30.0/51.7	35.0/66.1	36.2/68.3
Tainan	PBM Index	K-Means	***20.1/32.3***	21.4/36.9	21.0/33.3	22.6/35.9	21.8/32.7	23.9/36.8	21.9/34.8	24.3/37.2
		K-Medoids-mean	20.5/33.3	21.2/35.8	21.3/34.1	22.8/35.8	22.2/34.0	23.3/35.5	23.0/33.1	23.7/35.5
		K-Medoids-min	20.8/36.0	21.0/36.7	22.3/35.6	23.3/37.9	21.2/35.7	23.5/36.0	23.4/35.8	25.2/37.9
	DB Index	K-Means	21.3/33.6	21.3/36.7	21.0/34.1	22.4/36.5	22.0/33.4	22.9/36.5	22.3/34.8	24.6/38.3
		K-Medoids-mean	21.3/33.8	22.3/36.7	22.6/33.6	24.3/35.1	22.5/32.9	23.3/36.1	24.1/34.0	24.9/36.2
		K-Medoids-min	21.1/34.2	21.4/35.7	21.9/34.4	22.1/36.2	21.3/33.3	23.6/36.2	23.1/33.9	25.4/39.1

Table 6: The 50% and 80% MAPE results of forecasting daily electricity demand using the Tainan dataset with different features and cluster parameter settings

		FXF	FXT	TXF	TXT
PBM Index	K-Means	15.9/21.1	16.9/21.2	17.8/21.9	18.1/24.9
	K-Medoids-mean	***15.6/20.9***	16.0/21.2	16.9/21.3	18.4/24.3
	K-Medoids-min	15.7/22.2	16.9/22.7	17.7/21.5	18.7/24.2
DB Index	K-Means	18.2/24.9	19.8/30.1	19.3/27.6	19.5/29.9
	K-Medoids-mean	18.6/23.6	21.0/31.8	19.3/26.5	20.1/31.7
	K-Medoids-min	18.2/25.2	21.4/31.6	18.0/25.0	20.9/29.7

the floor number feature is considered (i.e., *FTF*). This is because there are many vacancies on the HIGH floors of properties in the Hsinchu dataset, so the floor number feature has a strong influence on the forecasts of electricity demand. Moreover, we observe that the 80% MAPE values are greater than 46.5% when the temperature feature is considered (i.e., *TFF*, *TFT*, *TTF*, and *TTT*). One of the reasons is that the correlation of the electricity demand and the temperature is between weak and moderate in the Hsinchu dataset, and the misuse of the temperature feature degrades the prediction accuracy significantly.

With regard to the Tainan dataset, the best forecast performance is achieved when the K-Means algorithm and the PBM index are used, but none of the three features are considered. The reasons are as follows.

1. Tainan has a tropical climate, which means the temperatures are relatively high throughout the year. Thus, the temperature feature has less impact on forecasts of electricity demand.

2. Unlike the other two datasets, the Tainan dataset shows the electricity demand for each floor of a three-floor house. Therefore, the per-floor electricity demand depends more on the lifestyles of the household's residents than the floor number (i.e., LOW, MIDDLE, or HIGH).

3. The Tainan site is located in a purely residential area, so holidays have less effect on daily electricity demands.

We also evaluated the forecast accuracy by considering the total household electricity demand in the Tainan dataset (i.e., the electricity demand of the three floors in a building). The evaluation results are shown in Table 6, where 'X' means "*don't-care*" of the corresponding feature. The best forecast performance occurs when the parameter settings are the same as those in the per-floor case; however, the 50%/80% MAPE values improve significantly in the per-household results (15.6/20.9 compared to 20.1/32.3 in the per-floor case). The results confirm that there exists a lifestyle factor in the per-floor-based dataset. Nevertheless, by aggregating the electricity demand of floors in the same building, it is possible to derive accurate forecasts of electricity demand in the Tainan scenario.

7. CONCLUDING REMARKS

We have conducted in-depth data analysis and forecast of household electricity demand using three realistic datasets of different household lifestyles. The analysis shows that household electricity demand is highly correlated to the temperature, the floor number, and the type of day in all the datasets at different scales. Moreover, using various parameter settings for data clustering and the SVR method, we evaluate the accuracy of forecasts of daily electricity demand in the three datasets. The results demonstrate that there exists a life style diversity between the three datasets, and the best forecast performance of each dataset is derived under different parameter settings. Specifically, the medium MAPE of the best forecast achieved is 15.6%. Research on finer-grained forecasts of household electricity demand is ongoing. We hope to report the results in the near future.

8. ACKNOWLEDGEMENTS

This research was partially supported by the National Science Council (NSC 101-2628-E-001-004-MY3) and the Bureau of Energy of the Ministry of Economic Affairs, Taiwan.

9. REFERENCES

[1] Electricity load forecast using intelligent adaptive technology. http://neuron-ai.tuke.sk/competition/, 2001.

[2] Assessment of demand response and advanced metering. FERC Staff Report, October 2013.

[3] AERA, APA, and NCME. *Standard for Educational and Psychological Testing*. American Educational Research Association, 1999.

[4] N. Atamturk, M. Zafar, and P. Clanon. Electricity use and income: A review. Technical report, Policy and Planning Division Literature Review, California Public Utilities Commission, June 2012.

[5] S. Bubeck, M. Meila, and U. von Luxburg. How the initialization affects the stability of the k-means algorithm. *ESAIM: Probability and Statistics*, 16:436–452, January 2012.

[6] E. Castillo, B. Guijarro, and A. Alonso. Electricity load forecast using functional networks. In *EUNITE Symposium - Competition on Electricity Load Forcast Using Intelligent Technologies*, 2001.

[7] C.-C. Chang and C.-J. Lin. LIBSVM - A Library for Support Vector Machines. http://www.csie.ntu.edu.tw/ cjlin/libsvm.

[8] B.-J. Chen, M.-W. Chang, and C.-J. Lin. Load forecasting using support vector Machines: a study on EUNITE competition 2001. *IEEE Transactions on Power Systems*, 19(4):1821–1830, November 2004.

[9] D. L. Davies and D. W. Bouldin. A Cluster Separation Measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-1(2):224 – 227, April 1979.

[10] D. Esp. Adaptive logic networks for east slovakian electrical load forecasting. In *EUNITE Symposium - Competition on Electricity Load Forcast Using Intelligent Technologies*, 2001.

[11] A. Jain and B. Satish. Clustering based Short Term Load Forecasting using Support Vector Machines. In *IEEE Bucharest Power Tech Conference*, 2009.

[12] W. Kowalczyk. Averaging and data enrichment: two approaches to electricity load forecasting. In *EUNITE Symposium - Competition on Electricity Load Forcast Using Intelligent Technologies*, 2001.

[13] J. B. Kruskal and M. Liberman. *Time Warps, String Edits, and Macromolecules - The Theory and Practice of Sequence Comparison*, chapter The Symmetric Time-Warping Problem: from Continuous to Discrete, pages 125–161. Reading. Addison-Wesley Publishing Co., Massachusetts, September 1983.

[14] A. Lewandowski, F. Sandner, and P. Protzel. Prediction of electricity load by modeling the temperature dependencies. In *EUNITE Symposium - Competition on Electricity Load Forcast Using Intelligent Technologies*, 2001.

[15] A. Lotfi. Application of Learning Fuzzy Inference Systems in Electricity Load Forecast. In *EUNITE Symposium - Competition on Electricity Load Forcast Using Intelligent Technologies*, 2001.

[16] J. MacQueen. Some methods for classification and analysis of multivariate observations. In *Berkeley Symposium on Mathematical Statistics and Probability*, 1967.

[17] F. Martinez-Alvarez, A. Troncoso, J. C. Riquelme, and J. S. AguilarRuiz. Energy Time Series Forecasting Based on Pattern Sequence Similarity. *IEEE Transactions on Knowledge and Data Engineering*, 23(8):1230–1243, August 2011.

[18] B. L. Miller and D. E. Goldberg. Genetic Algorithms, Tournament Selection, and the Effects of Noise. *Complex Systems*, 8:1930212, 1995.

[19] A. K. Mishra, D. E. Irwin, P. J. Shenoy, J. Kurose, and T. Zhu. Smartcharge: Cutting the electricity bill in smart homes with energy storage. In *ACM e-Energy*, 2012.

[20] P. Moscato. On Evolution, Search, Optimization, Genetic Algorithms and Martial Arts: Towards Memetic Algorithms. C3P Report 826. California Institute of Technology,, 1989.

[21] M. K. Pakhira, S. Bandyopadhyay, and U. Maulik. Validity index for crisp and fuzzy clusters. *Pattern Recognition*, 37(3):487–501, March 2004.

[22] E. Pelikan. Middle-Term Electric Load Forecasting by Time Series Decomposition. In *EUNITE Symposium - Competition on Electricity Load Forcast Using Intelligent Technologies*, 2001.

[23] R. Poli and W. B. Langdon. Genetic Programming with One-Point Crossover. In *Soft Computing in Engineering Design and Manufacturing*, 1997.

[24] A. P. Reynolds, G. Richards, and V. J. Rayward-Smith. The Application of K-medoids and PAM to the Clustering of Rules. In *International Conference on Intelligent Data Engineering and Automated Learning*, 2004.

[25] W. Shen, V. Babushkin, Z. Aung, and W. L. Woon. An ensemble model for day-ahead electricity demand time series forecasting. In *ACM e-Energy*, 2013.

[26] A. J. Smola and B. Scholkopf. A tutorial on support vector regression. *Statistics and Computing*, 14(3):199–222, August 2004.

[27] D. M. Solomon, R. L. Winter, A. G. Boulanger, R. N. Anderson, and L. L. Wu. Forecasting Energy Demand in Large Commercial Buildings Using Support Vector Machine Regression. Technical Report CUCS-040-11, Department of Computer Science, Columbia University, 2011.

[28] P.-N. Tan, M. Steinbach, and V. Kumar. *Introduction to Data Mining*. Addison-Wesley, 2006.

[29] V. N. Vapnik. *Statistical Learning Theory*. Wiley-Interscience, 1st edition, September 1998.