Japan Architectural Review Open Access

Architectural
Institute of Japan

Original Paper

Electricity load forecasting using clustering and ARIMA model for energy management in buildings

Bishnu Nepal, Motoi Yamaha, Aya Yokoe and Toshiya Yamaji

Chubu University, Kasugai, Aichi, Japan

Correspondence
Bishnu Nepal, Chubu University, 1200 Matsumoto-cho,
Kasugai, Aichi, Japan.
Email: npl.bishnu1@gmail.com

Funding Information
No Funding Information provided.

Received September 11, 2019; Accepted November 21, 2019

doi: 10.1002/2475-8876.12135

Abstract

Understanding the energy consumption patterns of buildings and investing efforts toward energy load reduction is important for optimizing resources and conserving energy in buildings. In this research, we proposed a forecasting method for the electricity load of university buildings using a hybrid model comprising a clustering technique and the autoregressive integrated moving average (ARIMA) model. The novel approach includes clustering data of an entire year, including the forecasting day using K-means clustering, and using the result to forecast the electricity peak load of university buildings. The combination of clustering and the ARIMA model has proved to increase the performance of forecasting rather than that using the ARIMA model alone. Forecasting electricity peak load with appreciable accuracy several hours before peak hours can provide the management authorities with sufficient time to design strategies for peak load reduction. This method can also be implemented in the demand response for reducing electricity bills by avoiding electricity usage during the high electricity rate hours.

Keywords
ARIMA model, clustering, electricity load forecasting, energy conservation, K-means algorithm

1. Introduction

Clustering is a process of partitioning data objects into groups or clusters so that the objects within a cluster are similar to one another and dissimilar from the objects in other clusters.[1] Cluster analysis is an unsupervised learning method that acts as a cornerstone in intelligent data analysis processes, and it is used for the exploration of interrelationships among a collection of patterns by organizing them into homogeneous clusters.[2] This method is used for identifying different types of energy consumption and has been applied to individual, industrial, and commercial customers or a large aggregation of residential customers.[3,4] The development of smart meters with frequent interval data and better precision has motivated more researches to apply clustering and machine learning techniques used for industrial and commercial customers to individual residential customers.[5–8] Obtaining a large amount of data is of no use unless it is changed into something useful. Analyzing data can provide relevant information that helps in numerous applications such as market analysis, fault detection, production control, science exploration, and energy consumption forecasting. Inventions of new electrical appliances and technologies have increased in the past few years and are expected to increase in the upcoming future. Short-term or long-term prediction of electricity load can help energy management teams in their respective sectors to design strategies for energy consumption reduction. Thus, forecasting electricity load and electricity management is essential for the conservation of energy.

Time series forecasting has gained popularity among researchers over the past few decades. Important areas of research of time series forecasting includes business, economics, engineering, medicine, social sciences, and politics. Autoregressive integrated moving average (ARIMA) model is most popular for stochastic time series models. Recently, other methods such as artificial neural network (ANN) and support vector machine (SVM) have gained popularity in time series forecasting.[9,10] After the Great East Japan earthquake, peak load reduction and power saving have become a concern. Many Japanese universities and commercial buildings have invested efforts toward energy conservation and reduction in CO_2 emissions. For energy management and demand response in buildings, accurate electricity load forecasting is essential. The purpose of this research is to support energy management tasks such as measuring demand response and ensuring peak load reduction in university buildings through the development of a forecasting model that improves the forecasting accuracy in comparison to existing forecasting methods.

In this paper, we proposed a hybrid model, which is a combination of clustering and the ARIMA model, to produce more accurate peak load forecasting of university buildings. Using the auto.arima function, the selection of the ARIMA model has been made automatic so, and therefore, if we have only time series data, the forecasting of the electricity load with good accuracy can be achieved without considering the building schedule, occurrence of holidays, occurrence of events, etc for forecasting.

1.1 Motivation

Electricity fee is charged according to the amount of electricity consumed and the contracted value of the peak demand. If the electrical load exceeds the contracted value of the peak demand, customers need to pay more on the electricity bill and thus they avoid exceeding this contracted value.

Figure 1 shows the electricity load on July 17, 19, 23, and 6th August of the academic year 2018 in the West Campus of Chubu University. On these days, the electricity load of the West Campus of Chubu University exceeded the demand electricity load represented by the dotted red line shown in Figure 1. This motivated us for the determination of some automatic technique that uses past data to forecast the electricity load with suitable accuracy. Chubu University is divided into two electricity contract zones, that is, the East Campus zone and the West Campus zone. In this paper, the electricity load data of only the East Campus zone is used for the analysis.

1.2 Literature review

Various machine learning algorithms for short-term and aggregate forecasting of residential electricity consumption for 1 hour and one day ahead residential electricity forecasting is evaluated in.[11] The forecasting accuracy is evaluated using evaluation metrics that are scale independent and robust to a value approaching zero: normalized root mean square error (NRMSE) and normalized mean absolute error (NMAE). Classifying the electricity consumption of different households into a predefined number of clusters, and summing the forecasted value of each cluster‘s aggregated electricity consumption produces better result than forecasting the electricity consumption of each household individually and then summing them. A data processing system to analyze energy consumption patterns in industrial parks, based on a cascade application of a self-organizing map and the K-means clustering algorithm is presented in.[12] The system is validated using real load data from an industrial park in Spain. The validation results show that the system adequately finds different behavior patterns that are meaningful and is capable of doing so without supervision and without any prior knowledge about the data. A density-based clustering technique, CHSFDP, is performed to discover the typical dynamics of electricity consumption and segment customers into different groups.[13] A time domain analysis and entropy are conducted on the result of dynamic clustering to identify the demand response potential of the customers of each group. K-means clustering for time series forecasting in R for predicting the electricity consumption in the residential and industrial sectors of Oman is applied in.[14] Forecasting results for each cluster are analyzed using TSl + RWD, TBATS, ARIMA, etc and the prediction accuracy is calculated using the Mean Absolute Error and Root Mean Squared Error. The TBATS model is found to provide a more accurate result. However, this research was carried with only data for five months, and thus, at least one-year data would be desirable for better forecasting. Dynamic clustering algorithm is based on different criteria such as season, electricity consumption, and the field where the electricity is consumed is used in.[15] After clustering, a Deep Belief Network (DBN) based on load demand forecasting is implemented for each cluster, and then, the multi forecast DBN is implemented to forecast long-term demand forecasting. Monthly energy consumption forecasting of the year 2005 is performed using the data for the year 2004. A comprehensive study of clustering methods for residential electricity demand profile and further applications focused on the creation of more accurate electricity forecast for residential customers is presented in.[16] The determination of an appropriate number of clusters using various distance measures and forecasting results for aggregate models for the

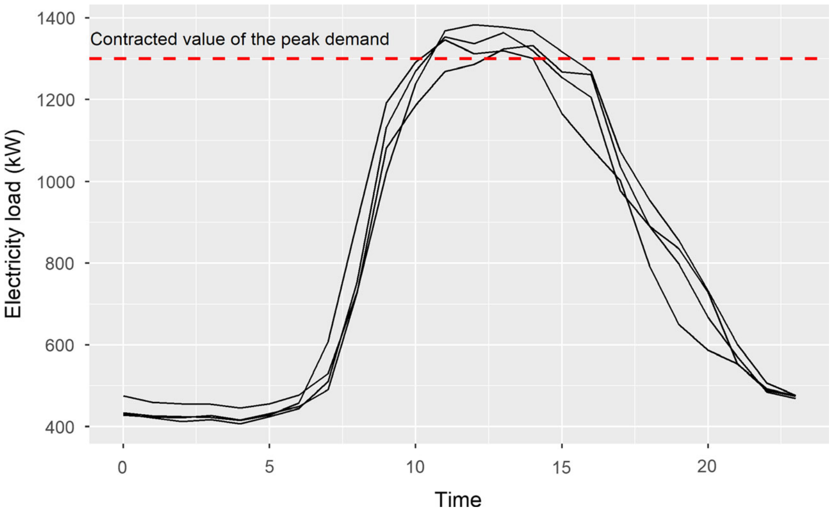


Figure 1. Electricity load graph with peak load crossing the contracted value of the peak demand in West Campus of Chubu University

Figure 2. Photograph of Chubu University Japan

entire population is discussed. A method to predict the electricity demand using linear regression (LR) and the support vector regression (SVR) is proposed in.[17] The SVR and LR are implemented using the weka tool that classifies and clusters the data. Error calculation shows that the proposed method provides improved accuracy as well as improved performances. Clustering for bottom-up short-term load forecasting is used in.[18] The disaggregation strategy uses a nonparametric model to handle forecasting and wavelets to define various notations of similarity between load curves, and it achieves a 16% improvement in forecasting accuracy when applied to French individual consumers.

1.3 ARIMA model

Box and Jenkins developed a mathematical model for forecasting a time series by fitting it with data and using the fitted model for forecasting (ie, ARIMA model).[19,20] In general, the ARIMA process is written with the notation ARIMA(*p*,*d*,*q*), where *p* denotes the number of autoregressive orders in the model. Autoregressive orders specify the previous values from the series which are used to predict current values; difference (*d*) specifies the order of differencing applied to the series before estimating models; and moving average (*q*) specifies how deviations from the series mean for previous values are used to predict current values.[21]

An ARMA consists of two parts, an autoregressive (AR) part and a moving (MA) part. The model is usually then referred to as the ARMA (*p, q*) model where *p* is the order of the autoregressive part and *q* is the order of the moving average part.[22,23]

$$X_t = c + \phi_1 X_{t-1} + \phi_2 X_{t-2} + \ldots + \phi_p X_{t-p} + \theta_1 \varepsilon_{t-1} + \cdots + \theta_q \varepsilon_{t-q} + \varepsilon_t \tag{1}$$

An ARIMA(*p*,*d*,*q*) model is a generalization of an ARMA model where *p*, *d*, and *q* are non-negative integers that refer to the order of the autoregressive, integrated, and moving parts of the model, respectively.

To deal with the seasonality of the ARIMA model, the generalized ARIMA model with the seasonal differencing of an appropriate order is used to remove non-stationary items from the series. For a monthly time series $s = 12$, and for quarterly time series $s = 4$. The model is generally termed as SARIMA $(p,d,q) \times (P,D,Q)^s$ model.

1.4 Target of the research

Figure 2 represents the aerial photograph of Chubu University, a private University in Japan located in Aichi prefecture. It has seven departments and around 10 000 students studying in both the science and the non-science departments.

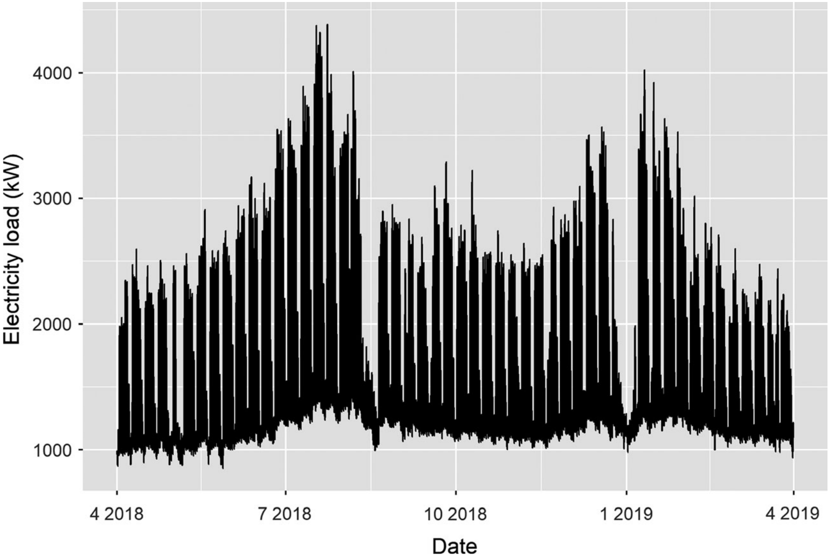


Figure 3. Line graph of whole university electricity load of academic year 2018

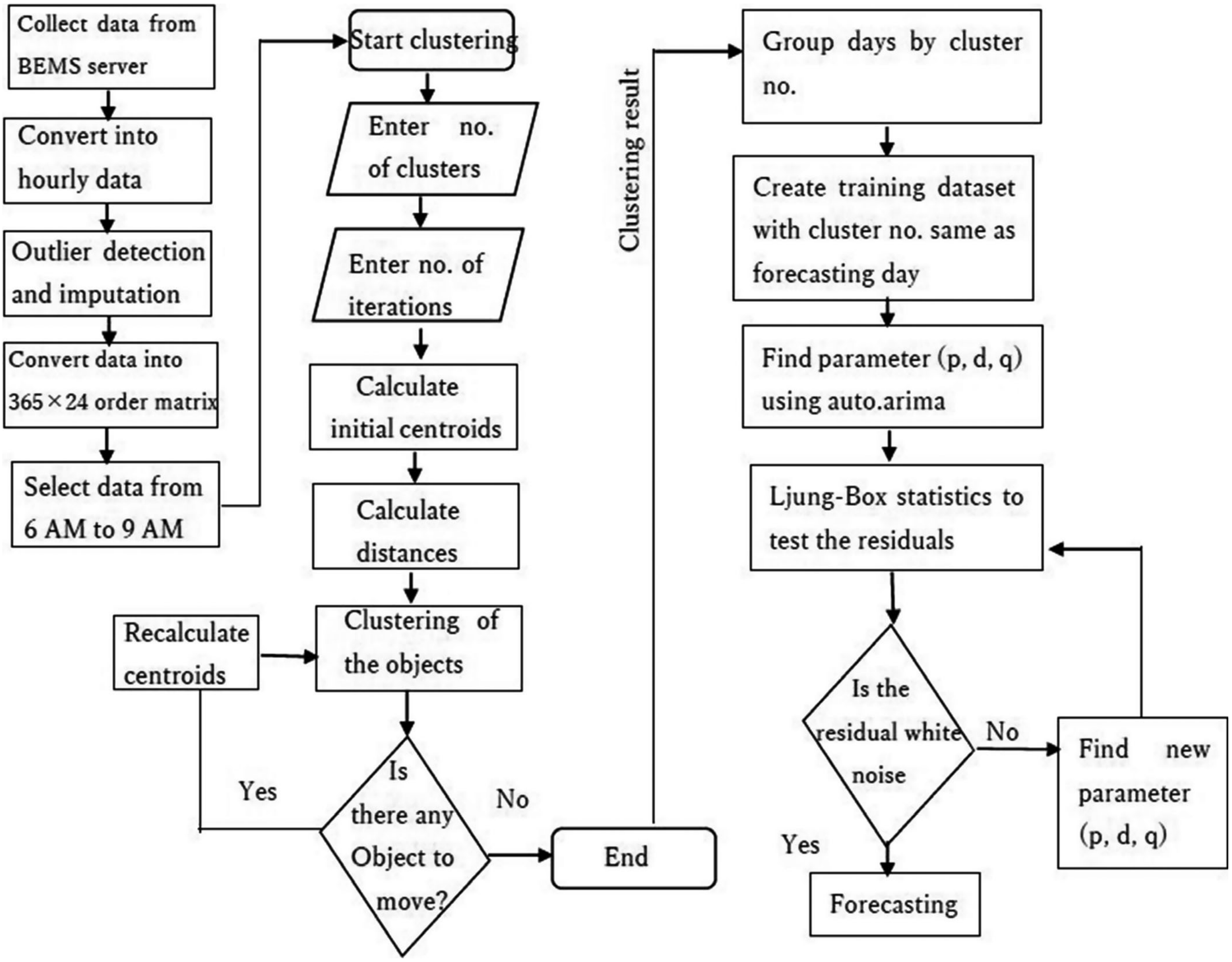

Collect data from BEMS server
Convert into hourly data
Outlier detection and imputation
Convert data into 365 × 24 order matrix
Select data from 6 AM to 9 AM
Start clustering
Enter no. of clusters
Enter no. of iterations
Calculate initial centroids
Calculate distances
Clustering of the objects
Recalculate centroids
Is there any Object to move?
Yes
No
End
Clustering result
Group days by cluster no.
Create training dataset with cluster no. same as forecasting day
Find parameter (p, d, q) using auto.arima
Ljung-Box statistics to test the residuals
Is the residual white noise
No
Find new parameter (p, d, q)
Yes
Forecasting

Figure 4. Flow chart of proposed method

As seen in Figure 3, the electric load of Chubu University changes over the year, with the highest electricity load in air-conditioning usage in the summer and winter seasons. During vacations, electricity is found to fall low, near to the base electricity load of the university. Throughout the week too, there is a rise and fall in the electricity load due to the effect of lectures, partial lectures (Saturdays), holidays (Sundays and public holidays), occurrences of events, etc.

2. Methodology

2.1 Outline of the proposed method

Data used in this research are the hourly electricity load data of the East Campus of Chubu University of the academic year 2017 and 2018. Chubu University has a Building Energy Management System (BEMS) where electricity load data are measured at the transformer of each building, and the data are collected in the BEMS server every minute. The minute data are summed to produce hourly data, and thus, one-year data consists of 8760 data for each building. Due to technical problems, data collected in the BEMS server contain outliers and missing values. The presence of outliers in the training data can lead to errors in the forecasting output. Thus, in this research the outliers present in the raw data are converted into not available (NA) values as in[24] and imputed with appropriate values using linear interpolation using the zoo package of R programming language.

Data free from missing values and outliers are then analyzed using the K-means clustering method. As in,[24] the number of cluster = 6 is found to appropriate and one whole year electricity load data are classified into six clusters. The main objective of this analysis is to forecast the peak energy consumption of university buildings and develop an automatic process that receives data from the BEMS sever and forecast the peak load so that some strategy can be implemented for peak load reduction. The electricity load in university buildings depends upon different factors such as presence of holidays, presence of lectures, use of air-conditioning, the occurrence of events, and use of experimental facilities. Thus, the energy load changes day-to-day. Forecasting of electricity load without considering the above factors can cause a high error in the forecasting result. Thus, to make an automatic system for data processing, we created a hybrid model with clustering and ARIMA model that clusters data so that the cluster

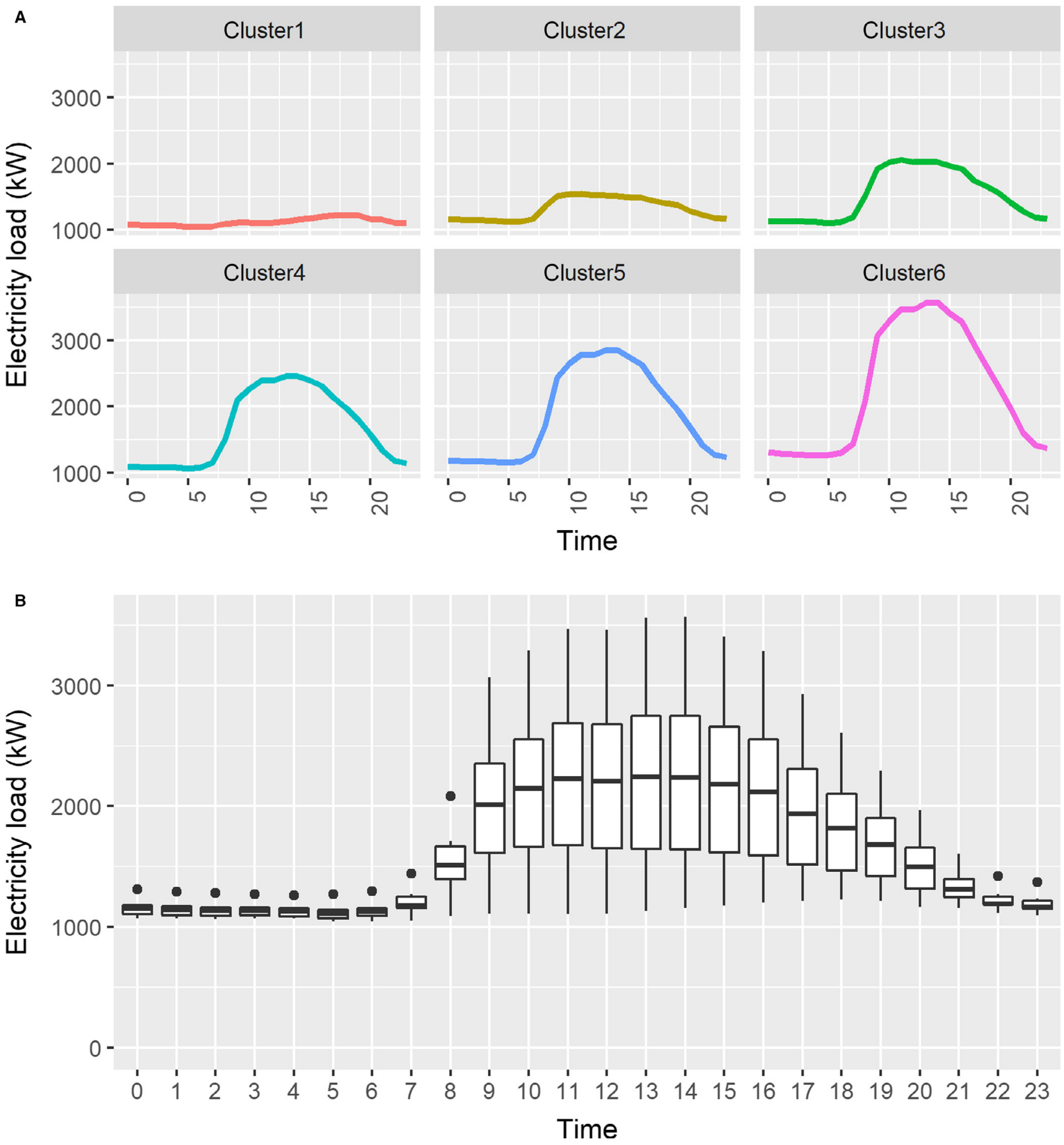


Figure 5. Clustering result of East Campus of Chubu University. A, Electricity load of cluster center 1 to 6. B, Electricity load distribution of cluster centers using box plot

number of the forecasting day is known. In the case of the ARIMA model, the selection is made automatically using the auto.arima function of R programming language.

The process from data collection to electricity load forecasting is shown in the flow chart in Figure 4. The data used in this paper are the building energy data of the East campus of Chubu University. The data collected from the BEMS server are 30-minute interval data converted into hourly data. After outlier detection and imputation of missing values, one-year data (365 days) including the forecasting day until 9 AM are created. These data are converted into a 365 × 24 order matrix, out of which electricity load data from 6 to 9 AM from all 365 days including the forecasting day until 9 AM are extracted, on which clustering is performed. The data from 6 to 9 AM of the forecasting day is used only for clustering, and not for electricity load forecasting. In this study, six clusters

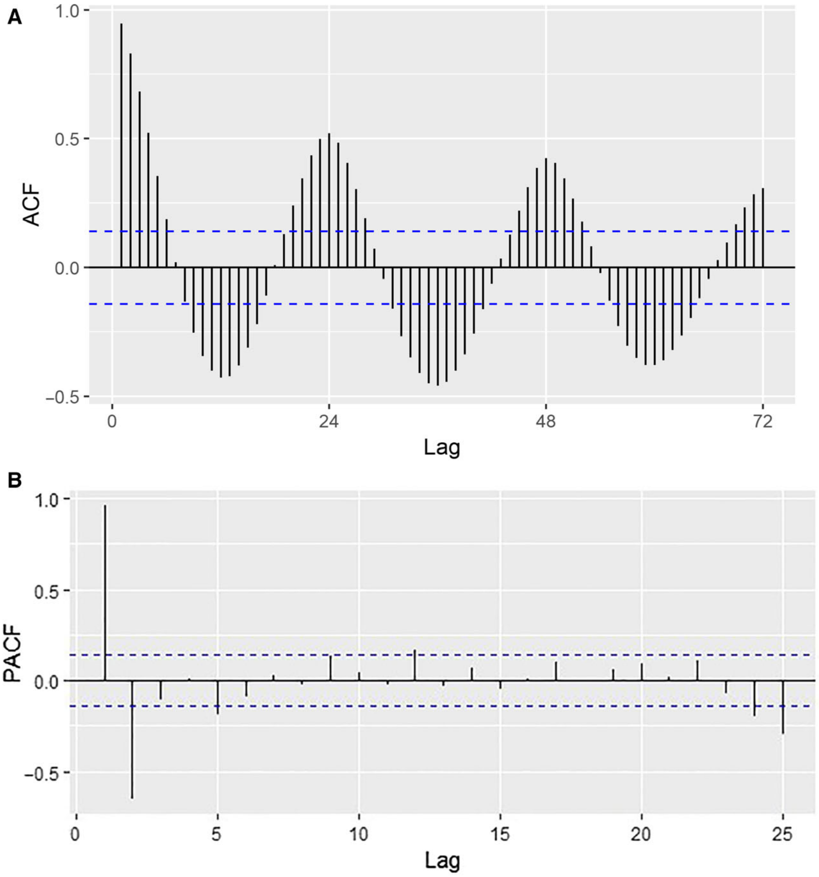


Figure 6. Autocorrelation function (ACF) (A) and partial autocorrelation function (PACF) (B) of the fitted model

Table 1. AIC values of suggested ARIMA models

ARIMA(*p*,*d*,*q*)(*P*,*D*,*Q*)$_S$	AIC
ARIMA(2,0,0)(0,0,0)$_{24}$	2105.4
ARIMA(2,0,1)(2,0,0)$_{24}$	2104.5
ARIMA(2,0,1)(2,0,1)$_{24}$	2096.4
ARIMA(2,0,2)(2,0,1)$_{24}$	2098.6
ARIMA(2,1,1)(2,0,1)$_{24}$	2097.6
ARIMA(3,0,1)(2,0,0)$_{24}$	2105.6
ARIMA(3,0,1)(2,0,1)$_{24}$	2097.8
ARIMA(3,0,1)(3,0,1)$_{24}$	2107.9

AIC, Akaike information criteria; ARIMA, autoregressive integrated moving average.

were found to suitable and the initial centroids are calculated using the percentile method. The clustering result classifies all 365 days into 6 clusters with days that have the similar electricity load characteristics. Then, a dataset with days belonging to the same cluster as the forecasting day is created, which is used as the training data for the forecasting model. The parameters (*p*, *d*, *q*) of the best model for the training dataset is calculated using the auto.arima function of R programming language. The next step is the testing of the residuals of the model using Ljung-Box statistics; if the residuals are a white noise, then the model is ready for forecasting, otherwise a new model needs to be selected.

The combination of clustering and the ARIMA model have been used before as discussed in the literature review section, but most of them are related to clustering group of customers and finding the aggregate forecasting of individual clusters. This concept of clustering the forecasting day used in this paper has never been discussed in any other journal before.

2.2 Clustering and accuracy measures

Using K-means clustering, we classified the electricity load of the university into six clusters based on the hourly distribution of one-year electricity load.

Figure 5A is the electricity load plot of each cluster and Figure 5B is the box plot showing the electricity load distribution of Chubu University per cluster and hours of the day respectively. A large difference in the energy load from cluster 1 to cluster 6 is observed and even the base energy is found to increase with a rise in the cluster number. The electric load from 10 PM till 6 AM is almost similar. From 8 AM, the electric load is found to rise sharply until it reaches its peak around 1 or 2 PM. As seen in Figure 5A, the electricity peak load for each of the cluster is different, and thus, by clustering the forecasting day and using its cluster members for electricity load prediction can improve the peak load forecasting accuracy of any forecasting algorithm.

Many comparative studies have been conducted with the aim of identifying the most accurate methods for time series forecasting.[25] However, research findings indicate that the performance of forecasting methods varies according to the accuracy measure being used. A good accuracy measure should provide information on the clear summary of the error distribution. Root mean square error (RMSE), mean absolute

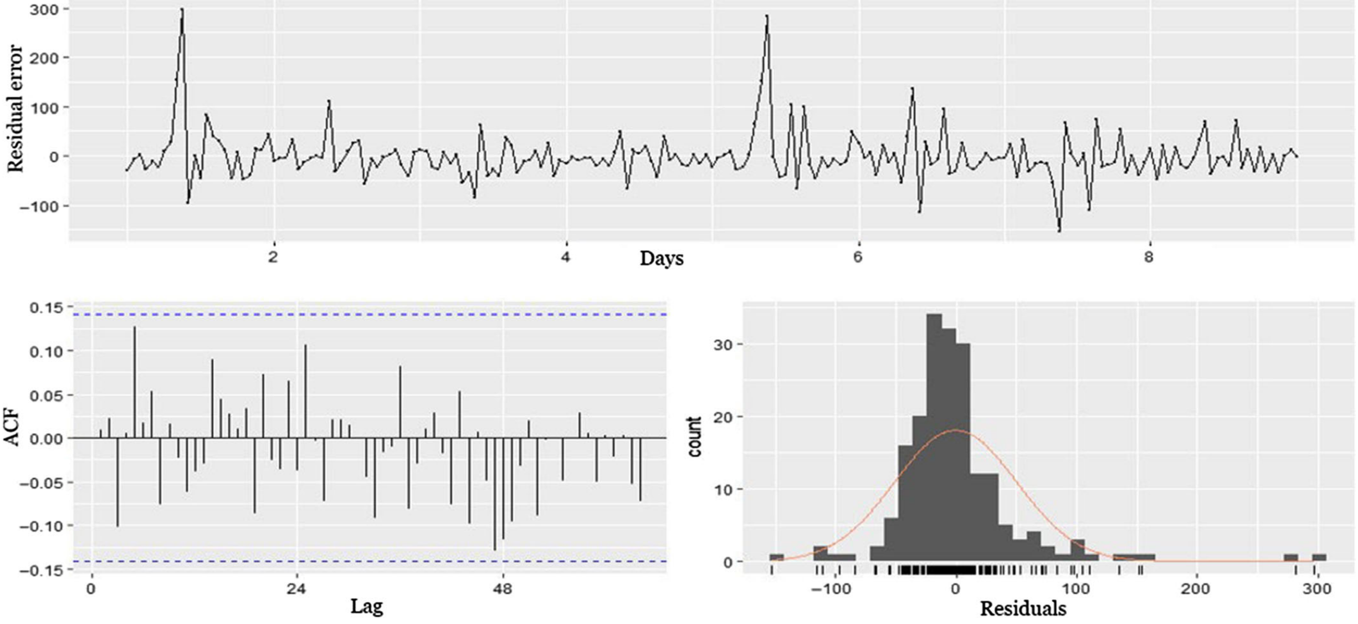


Figure 7. Residual analysis for the autoregressive integrated moving average (ARIMA)(2,0,1)(2,0,1)$_{24}$ fit to the data

percentage error (MAPE), and mean absolute error (MAE) are the very early and most popular accuracy measures. In this paper, we use these accuracy measures for the forecasting results.

RMSE and MAE are scale-dependent measures since their values depend on the scale of the data. They are useful in comparing forecasting methods on the same data. For 24 hours ahead forecasting, if "e_t" represents the error in forecasting for each hour, RMASE and MAE can be defined as

$$\mathrm{MAE} = \frac{1}{n}\sum_{t=1}^{n} |e_t| \tag{2}$$

$$\mathrm{MSE} = \frac{1}{n}\sum_{t=1}^{n} e_t^2 \text{ and } \mathrm{RMSE} = \sqrt{\mathrm{MSE}} \tag{3}$$

MAPE is based on percentage error of the observation and is scale-independent.

$$\mathrm{MAPE} = \frac{1}{n}\sum_{t=1}^{n} \frac{|e_t|}{|Y_t|} \tag{4}$$

2.3 Fitting the ARIMA model

The autocorrelation function (ACF) and partial autocorrelation function (PACF) are used to estimate the value of parameter p and q in the ARIMA model. Figure 6 shows the ACF and PACF of electricity load data of the training dataset for electricity load forecasting on 2nd April 2018. Based on the first two lags, it might be possible that AR (2) works based on the first two spikes in the PACF. To identify stationary/non-stationary processes of the time-series, we use augmented Dickey–Fuller test (ADF). The null-hypothesis of an ADF test is that the data is non-stationary, and small p-values suggest stationary. On using the ADF test using the adf.test function of R programming language, the series was found stationary. Thus, the parameter d in the ARIMA model can be chosen to be zero.

To find the best model for this data, we fit different models and select the model with the minimum Akaike information criteria (AIC) value. AIC is an estimator of the relative quality of statistical models for a given data set. The selection of the model is important, as under-fitting a model may not capture the true nature of the variability in the outcome variable, an over-fitted model loses generality. AIC is then a way to select the models that best balances these drawbacks and a smaller value of AIC represents a better model.[26]

In Table 1, the smallest value of AIC is obtained for seasonal ARIMA with the non-seasonal part (p,d,q) = (2,0,1) and the seasonal part of the model (P,D,Q) = (2,0,1). Then, we calculated the parameters of the best model for the training data using the auto.arima function of R programming language. The best model produced by auto.arima was "ARIMA (2,0,1)(2,0,1)$_{24}$" with AIC value 2096.4, same as the least value of AIC in the table.

Figure 7 shows the plot of residuals, its ACF values, and the histogram. As the ACF of residuals are within the significance level, the model selected using auto.arima is best fit for the training dataset. As the days pass on new data are added into the system and these data will be used in the forecasting in the future days. Thus, the training data used for forecasting goes on changing with time. Appropriate selection of these parameters (p, d, q) can only provide accurate forecasting results. Thus, we decided to select the best model using auto.arima functions for other training datasets. However, we use the Ljung–Box test for confirming whether the residuals are white noise.

Some other reasons for making the forecasting automatic includes

1. Fitting a model is not a simple task and most users are not expert at fitting time series models.
2. Automatic algorithms can produce better models than many experts.
3. Many businesses need hundreds of forecasts per week/month and they need it fast.

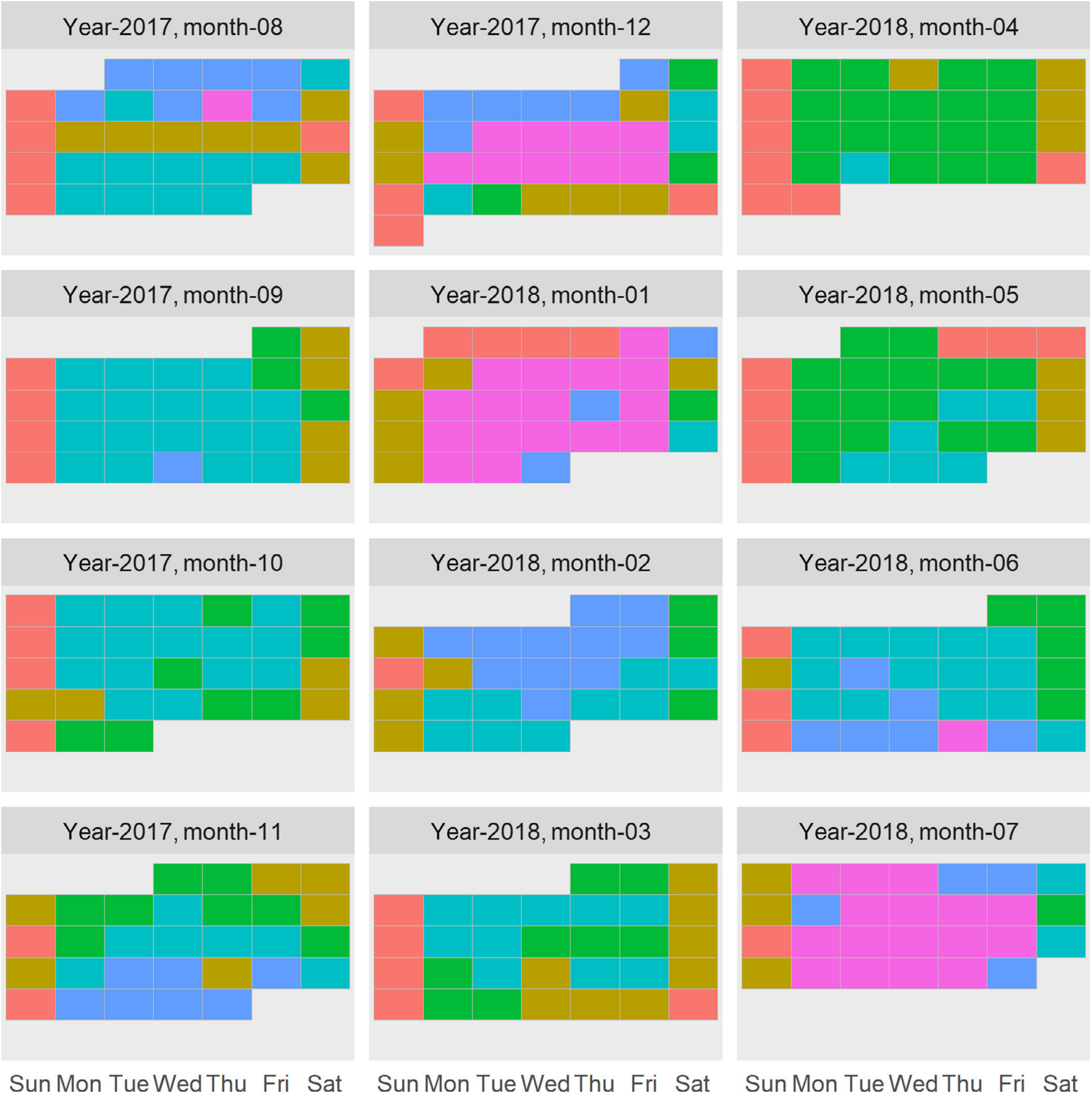


Figure 8. Calendar plot of clustering result with number of clusters K = 6

4 Some multivariate forecasting methods depends on many univariate forecasts.

3. Results

To carry out the proposed algorithm, it is necessary to determine the cluster number of the forecasting day. For this purpose, it is essential to select the dataset according to the forecasting day. In this research, data for two years from April 1, 2017 to the March 31, 2019 are used. If the forecasting of electricity load on March 1, 2019 is expected, then one-year data starting from March 2, 2018 to March 1, 2019 is created. The electricity load data from 6 to 9 AM of whole one year including the forecasting day are extracted. Then, K-means clustering is performed on the extracted data that is converted into a 365 × 4 order matrix by extracting 6 to 9 AM data from 365 days. The initial centroids required for K-means clustering are determined using the percentile method.[27] The entire one-year data are classified into six clusters; and each cluster representing the group of days has similar electricity load characteristics.

Calendar plot of Figure 8 is the clustering result of the one-year dataset including the forecasting day. The forecasting day, July 27, 2018 is assigned to cluster 5. In the case of

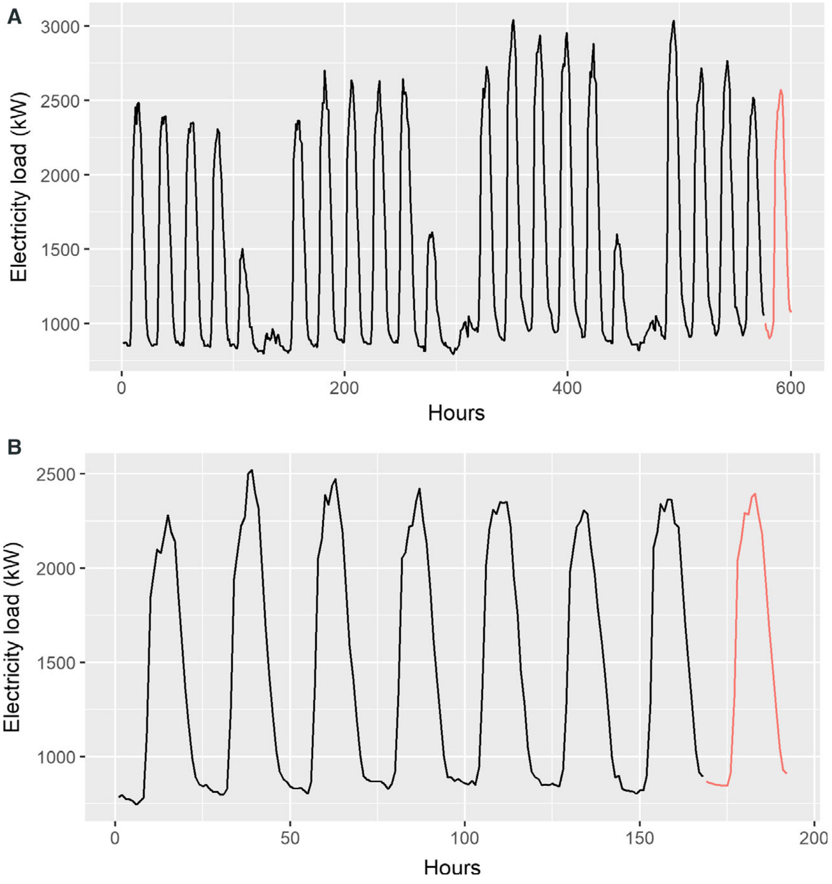


Figure 9. One day ahead forecasting of electricity load using autoregressive integrated moving average (ARIMA) model and proposed method. A, Forecasting result ARIMA model. B, Forecasting result proposed method

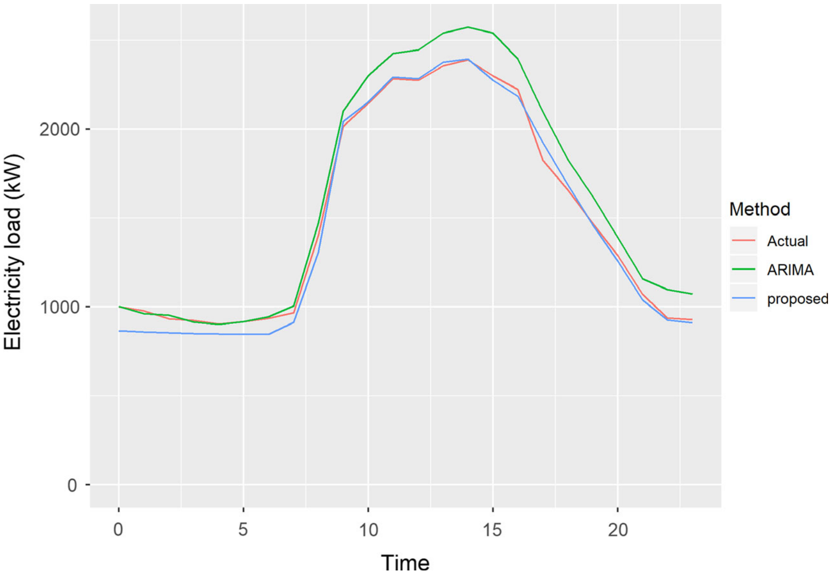


Figure 10. Forecasting result of autoregressive integrated moving average (ARIMA) and proposed method in comparison to actual data on July 27, 2018

ARIMA forecasting, the training data is chosen a few days before the forecasting days. Thus, a variety of days with different electricity load pattern are mixed in the training data, which are used to forecast the electricity load of the next day, which may produce forecasting with high errors. In the case of the proposed method, the training data are selected using the

Table 2. Forecasting result suggested ARIMA model and proposed method

Time (t)	Actual data (A_t)	Forecast using ARIMA (F_A)	Forecast using proposed method (F_p)	Error ARIMA model (E_A)	Error proposed method (E_p)
0:00	1001	1003.2	866.8	−2.2	134.2
1:00	979	962.8	859.6	16.2	119.4
2:00	933	953.0	855.6	−20.0	77.4
3:00	924	915.8	851.2	8.2	72.8
4:00	905	901.6	848.7	3.4	56.3
5:00	917	916.7	846.4	0.3	70.6
6:00	936	944.2	845.6	−8.2	90.4
7:00	967	1006.5	913.9	−39.5	53.1
8:00	1404	1473.2	1310.2	−69.2	93.8
9:00	2016	2102.6	2044.0	−86.6	−28.0
10:00	2148	2302.0	2155.2	−154.0	−7.2
11:00	2285	2426.8	2293.5	−141.8	−8.5
12:00	2275	2446.9	2283.6	−171.9	−8.6
13:00	2357	2541.3	2376.6	−184.3	−19.6
14:00	2390	2574.1	2394.8	−184.1	−4.8
15:00	2300	2541.2	2274.9	−241.2	25.1
16:00	2224	2394.8	2184.6	−170.8	39.4
17:00	1827	2100.8	1926.1	−273.8	−99.1
18:00	1658	1830.4	1688.6	−172.4	−30.6
19:00	1471	1622.4	1461.7	−151.4	9.3
20:00	1289	1391.0	1258.2	−102.0	30.8
21:00	1071	1159.3	1039.9	−88.3	31.1
22:00	938	1098.3	927.1	−160.3	10.9
23:00	930	1073.5	910.1	−143.5	19.9
MAE		108	47.5		
RMSE		134.6	60.7		
MAPE		6.6	4.2		

ARIMA, autoregressive integrated moving average; MAE, mean absolute error; MAPE, mean absolute percentage error; RMSE, root mean square error.

Table 3. Error values of forecasting result by ARIMA model and proposed method

	ARIMA model			Proposed method		
	MAPE	RMSE	MAE	MAPE	RMSE	MAE
SUN	32.9	364	261.2	5.8	61.9	47.6
MON	18.4	413.1	299.9	5.2	90.4	68.1
TUE	11.9	258	188	4.3	83.8	60.7
WED	9.9	213.4	155.8	4.3	78.7	59.5
THU	5.8	99.4	73.2	3.8	69.0	51.4
FRI	5.8	100.9	75.6	5.0	86.2	65.2
SAT	27.9	406.8	296.7	7.6	102.4	77.4
One year	16.1	265.3	193.1	5.1	81.7	61.3

ARIMA, autoregressive integrated moving average; MAE, mean absolute error; MAPE, mean absolute percentage error; RMSE, root mean square error.

clustering method. This makes the member of the training data have similar electricity load characteristics as the forecasting day and hence increasing the accuracy of the forecasting result.

Figure 9 is the one day ahead forecasting of electricity load. Figure 9A,B are the forecasting results of the ARIMA model and proposed method, respectively, on 27 July 2018. Black lines represent the training datasets whereas the forecasting result is indicated by red color. 24 preceding days are used as the training dataset in the ARIMA model whereas a group of days belonging to the same cluster as the forecasting day is used in the proposed method. The training dataset for ARIMA model consists of 24 days (576 hours), whereas the training data for the proposed method consists of 7 days (168 hours). The best model for both ARIMA and proposed method is calculated using the auto.arima function in R programming language. The computational time of the ARIMA method and the proposed method was measured using system.-time function of the R programing language using computer with a processor, Intel(R) Core(TM)i7-6700 CPU @3.4 GHz and 16 GB RAM. The elapsed time for the ARIMA model was 9.30 seconds whereas the elapsed time for the proposed method was significantly lower at 0.45 seconds.

The comparison between forecasting result of ARIMA model and proposed method using East campus data is shown in Figure 10.

From Table 2, the values of MAE, RMSE, and MAPE of the proposed method are found to be small in comparison to that of the ARIMA model. Results from Figure 10 and the values of MAE, RMSE, and MAPE of the proposed method shows that the proposed method produces better results than the ARIMA model.

The electricity load forecasting of each day using the ARIMA model and the proposed method from April 1, 2018 till March 31, 2019 was performed. The MAPE, RMSE, and MAE of each day between the actual data and the forecasted data using the ARIMA model and proposed model were calculated. Table 3 shows the one-year average value of MAPE, RMSE, and MAE between the actual and the forecasted value using the ARIMA model and the proposed method by day of the week.

Figure 11 is the average value of the accuracy test between ARIMA model and the proposed method. Figure 11A-C represent the bar graph showing the values of MAPE, RMSE, and MAE for the ARIMA and proposed method, respectively. The values of MAPE, RMSE, and MAE for ARIMA model is quite high at Saturday, Sunday, Monday, and Tuesday in comparison to the proposed method. As ARIMA model forecasts based on the electricity load of the previous days, sudden rise and fall of electricity load before the forecasting days which can occur due to the holiday (Sunday), partial holiday (Saturday), and the occurrence of events like campus festivals, open campus, sports events etc can affect on the forecasting accuracy.

Figure 12 is the one day ahead forecasting result after removing Saturday and Sunday from the dataset. Figure 12A, B shows the forecasting results of ARIMA model and proposed method respectively on September 18, 2018. To remove the effect of the holiday (Sunday) and partial holiday (Saturday), a dataset was created removing all Sunday and Saturday data from the academic year data 2018. Although there is no presence of Saturday and Sundays on this data, fluctuations in the electricity load data can be found. The presence of reduced electricity load on September 17 (a day before forecasting) led to a considerable error in the forecasting result on September 18 as shown in Figure 13. In the case of the proposed method, the presence of increased or reduced electricity load a day or few days before the forecasting day does not affect the forecasting result because the training data of the proposed method are the days belonging to the same cluster group with similar electricity load characteristics as the forecasting day.

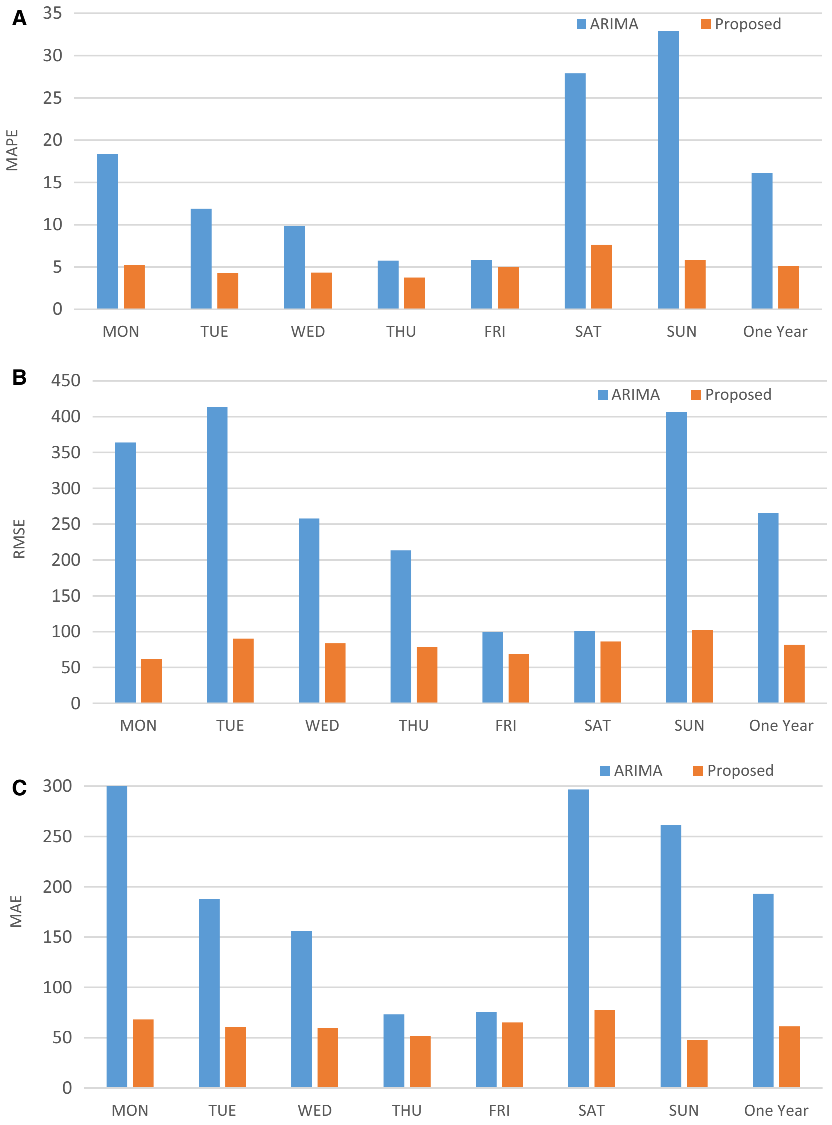


Figure 11. Bar graph showing the error values between autoregressive integrated moving average (ARIMA) and proposed method. A, Mean absolute percentage error (MAPE) value for ARIMA and proposed method. B, Root mean square error (RMSE) value for ARIMA and proposed method. C, Mean absolute error (MAE) value for ARIMA and proposed method

The MAPE, RMSE, and MAE values between the actual and forecasted values when Saturday and Sunday are removed using the ARIMA model are 18.5, 311.5, and 253.7, respectively. Whereas, the MAPE, RMSE, and MAE values between the actual and forecasted values using the proposed method is significantly lower at 2.7, 50.5, and 36.4, respectively.

Grouping by the day of the week can be a slightly better option than using only random days for forecasting. For example, to forecast the energy load on January 18, 2018 (Thursday), the training data for ARIMA model can be constructed by collecting previous days belonging to the same day of the week (Thursday), as energy load is likely to be similar on the same day of the week, rather than another day of the week. In Figure 8, the forecasting day, January 18, 2018 belongs to cluster 5. The days belonging to the same day of the week (Thursday) on the preceding weeks are found to belong to clusters 6, 2, and 1. [Correction added on 28 December 2019, after first online publication: The forecasting day for the energy load has been corrected to January 18, 2018 (Thursday) in this paragraph.] This will cause deviation in the forecasting values in the forecasting day. Exclusion of holidays and grouping by the days of the week can improve the accuracy of forecasting when done manually; however, the accuracy is less in comparison to the proposed model because the exclusion of holidays and grouping by the day of the week cannot ensure that the training data have similar characteristics as in the case of clustering. Thus, clustering is the best method to group the days on the basis of their similar electricity load characteristics.

4. Application of the Proposed Method

For peak load reduction and conservation of energy in Chubu University, we developed a visualization and analysis platform using Shiny App. Shiny App is a package of R programming language that makes it easy to build interactive web applications straight from R.

Some function developed in the Shiny App includes clustering and its resulting visualization in the form of cluster centers and a calendar plot, which allows viewing the any range of

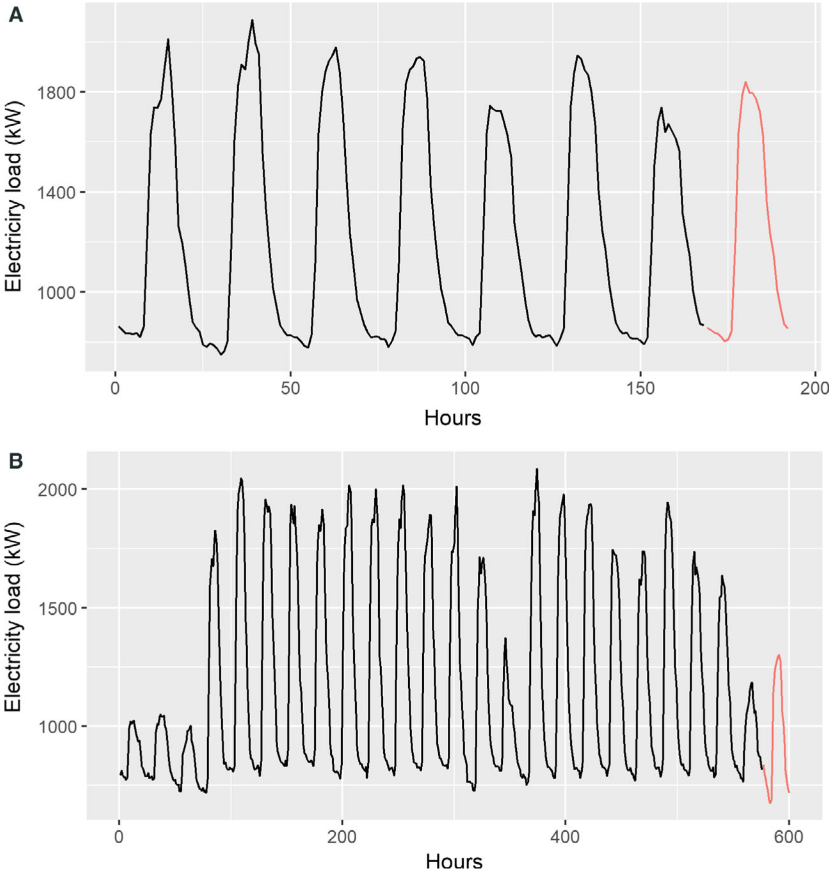


Figure 12. One day ahead forecasting of electricity load with dataset removing Saturday and Sunday. A, Forecasting using autoregressive integrated moving average (ARIMA) model. B, Forecasting using proposed model

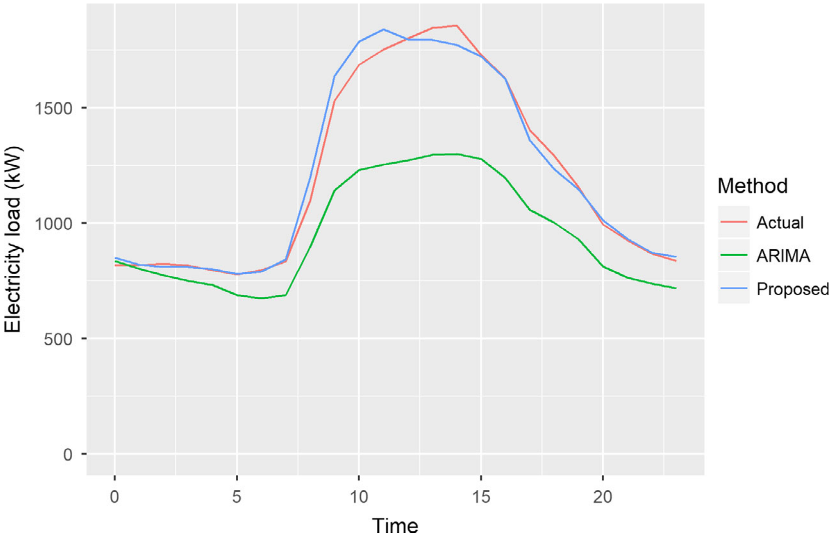
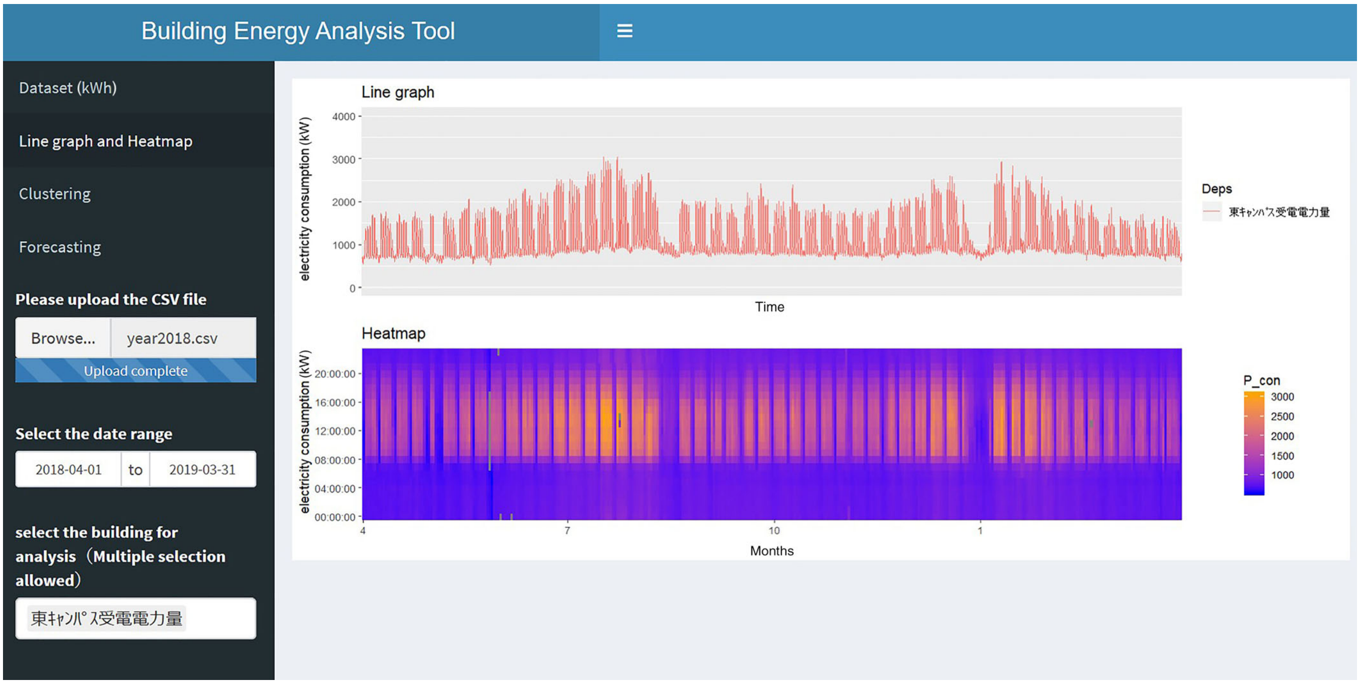


Figure 13. Forecasting result of autoregressive integrated moving average (ARIMA) and proposed method removing Saturday and Sunday on the training data

electricity load data available in the Shiny App in the form of trend graph and heat map. The selection of the building menu is also included that allows the user to analyze and view the electricity load characteristics of the desired building. Figure 14 is the visualization of one-year data in the form of a trend graph and heat map. The proposed forecasting model using Shiny App is as shown in Figure 15, and in the figure, it shows the comparison between the forecasting result on July 9, 2018 using the proposed method (light blue color) and the actual value (red color) until 2 PM.

Once the app is deployed on a web-based server such as Amazon Web Service (AWS) anybody with the server address



Building Energy Analysis Tool

Dataset (kWh)

Line graph and Heatmap

Clustering

Forecasting

Please upload the CSV file

Browse... year2018.csv

Upload complete

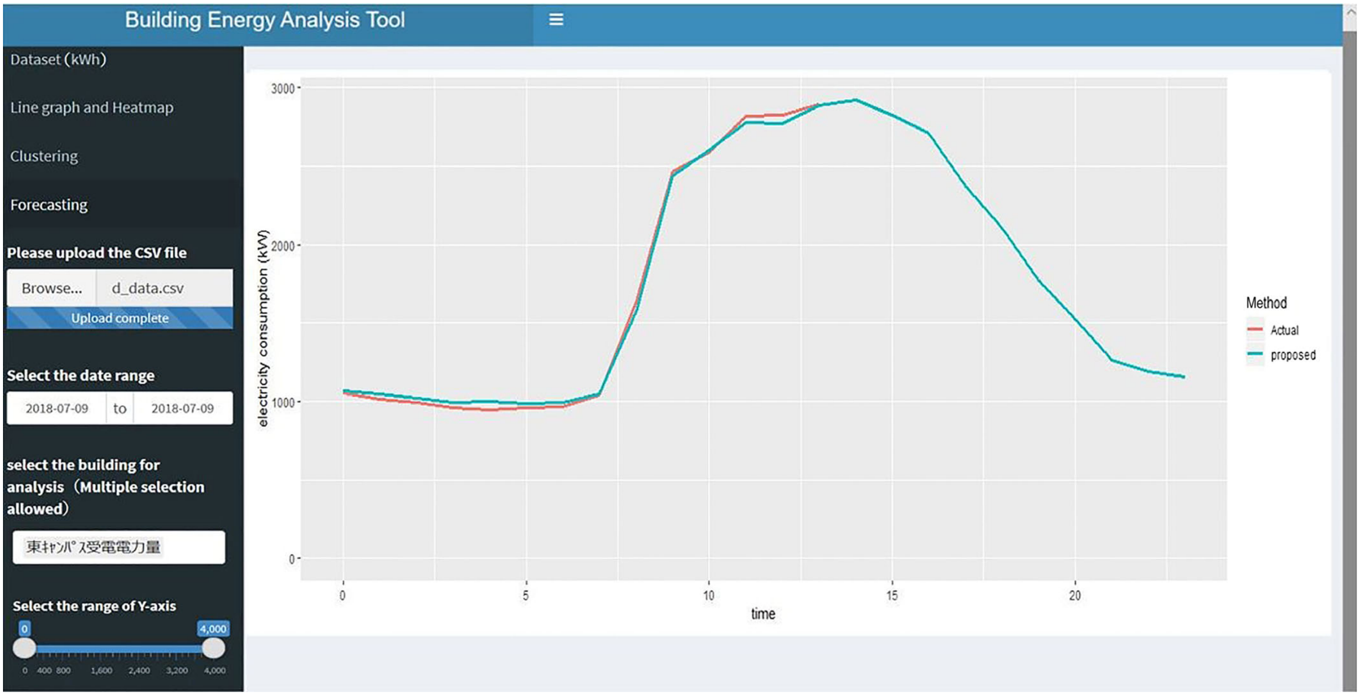
Select the date range

2018-04-01 to 2019-03-31

select the building for analysis（Multiple selection allowed）

東ｷｬﾝﾊﾟｽ受電電力量

Figure 14. Line graph and heat map of one-year data using the Shiny App



Building Energy Analysis Tool

Dataset (kWh)

Line graph and Heatmap

Clustering

Forecasting

Please upload the CSV file

Browse... d_data.csv

Upload complete

Select the date range

2018-07-09 to 2018-07-09

select the building for analysis（Multiple selection allowed）

東ｷｬﾝﾊﾟｽ受電電力量

Select the range of Y-axis

0 4,000

Figure 15. Forecasting result of electricity load on 9th July, 2018 using the Shiny App

can access the app and interact with it. The real-time data can be uploaded into the server every hour and the real-time comparison between the forecasted and the actual values can be performed. Thus, the efforts taken for peak load reduction can also be seen in real time.

The proposed method using Shiny App and AWS includes

- Making a group e-mail and SNS of energy conservation members of Chubu University which includes professors, master's course students, and faculty staff.

- Sending a mail with the link to the AWS server to the group members at 9:10 AM on weekdays.
- Requesting the group members to act toward peak load reduction based on the forecasted result.

Deploying the Shiny App to the web-based server has made it possible to view the energy status not only on the computer but also on the smartphone, and this makes the Shiny App more accessible regardless of the location of group members. Moreover, the selection of buildings makes it possible to view the energy pattern of the buildings in which they are interested. Hence, through the proper usage of the Shiny App and cooperation of the group members, electricity peak load reduction can be contributed.

5. Discussion

In this paper, we proposed a method to increase the performance of the ARIMA model using the clustering technique and forecasting the electricity load automatically. Using the proposed method, the analyst does not need to think about the energy load pattern on the previous days (eg, Holidays, occurrences of events) as the training data in the proposed method are created using the group members belonging to the same cluster as the forecasting day.

The proposed method can be useful for the conservation of energy in buildings and demand response. The electricity power company usually applies higher rates to their customers in the day time than in the night. Higher consumption of electricity in the electricity peak hours in the day can lead to higher electricity bills. Forecasting electricity load before reaching the peak electricity load can provide management authorities, faculty staff, and students with sufficient time to design a strategy for reduction of peak electricity load. For example, the authorities can decide the time of discharging the battery, time for using cogeneration, etc before the peak hours and faculty staff and students can take actions for reducing electricity consumption if they are informed about the forecasting. Moreover, this method can be useful in the demand response for reducing electricity bills by avoiding electricity usage during the high electricity rate hours. Based on the forecasting result and the using of the Shiny App, we proposed a method for peak load reduction. Every day (excluding weekends), a mail is sent to the group members of the energy conservation team with the status of the forecasted electricity load of the day with the link address of the Shiny AWS. The members can interactively view the energy load pattern of the forecasted day of the desired building and act accordingly for peak load reduction.

6. Conclusion

In this paper, we analyzed the electricity load pattern of Chubu University. Forecasting electricity load is important for energy conservation because it provides the energy conservation team with sufficient time and information for making a strategy and to actively implement it to achieve electricity load reduction. In this paper, a hybrid model of clustering and ARIMA model is used for improving the accuracy of the forecasting result and to automate the forecasting of electricity load. For clustering the electricity use data, one-year data including the forecasting day from 6 to 9 AM is used. Then, the one-year data are classified into six clusters using the K-means clustering algorithm. The members of the day of the forecasting are used as the training data for forecasting the electricity load. The probability that the cluster members of the forecasting day have an electricity load pattern similar to the forecasting is quite high in comparison to other clusters. This novel concept of collecting days with similar electricity load patterns for forecasting on a particular day has proved to produce better result than the ARIMA model. The accuracy of the forecasted result of the proposed method and ARIMA model are compared by calculating the MAPE, RMSE, and MAE values. One-year forecasting result from April 1, 2018 to March 31, 2019 shows that the values of MAPE, RMSE, and MAE of the ARIMA model is 16.1, 265.3, and 193.1, whereas the proposed method is significantly lower at 5.1, 81.7, and 61.3, respectively. The proposed method can be useful for electricity load reduction in buildings. Forecasting electricity load before reaching the peak electricity load can provide management authorities with sufficient time for making a strategy for peak electricity load reduction. The energy conservation group members viewing the forecasted value of the electric load of different buildings using AWS, and sending a request mail to act for peak load reduction depending on the status of the forecasted value are believed to help in the peak load reduction of University buildings.

Acknowledgement

The authors extend their appreciation to the Facility Administration Department of Chubu University for providing data and advice regarding building energy analysis.

Disclosure

The authors have no conflict of interest.

References

1 Han J, Kamber M, Pie J. *Data Mining Concepts and Technique*. Waltham, MA: Academic Press; Morgan Kaufmann Publisher; 2012.

2 Kotsiantis S, Pintelas PE. Recent advances in clustering: A brief survey. *WSEAS Transactions on Information Science and Applications*. 2004; **1**(1):73–81.

3 Chicco G, Napoli R, Piglione F. Comparison among clustering techniques for electricity customer classification. *IEEE Trans Power Syst*. 2006; **21**(2):933–940.

4 Chicco G. Overview and performance assessment of the clustering methods for electrical load pattern grouping. *Energy*. 2012;**42**(1):68–80.

5 Haben S, Singleton C, Grindro P. Analysis and clustering of residential customers energy behavioral demand using smart meter data. *IEEE Trans Smart Grid*. 2016;**7**(1):136–144.

6 Coa HÂ, Beckel C, Staake T. Are domestic load profiles stable over time? An attempt to identify target households for demand side management campaigns. *IECON 2013–39th Annual Conference of the IEEE Industrial Electronics Society*. 2013;4733–4738.

7 McLoughlin F, Duffy A, Conlon M. A clustering approach to domestic electricity load profile characterization using smart metering data. *Appl Energy*. 2015;**141**:190–199.

8 Rhodes JD, Cole WJ, Upshaw CR, Edgar TF, Webber ME. Clustering analysis of residential electricity demand profiles. *Appl Energy*. 2014;**125**:461–471.

9 Kohiro JM, Otienio RO, Wafula C. Seasonal time series forecasting: a comparative study of ARIMA and ANN models. *Af J Sci Technol*. 2004;**5**(2): 41–49.

10 Pan X, Lee B. A Comparison of support vector machines and artificial neural networks for mid-term load forecasting. *IEEE International Conference on Industrial Technology*. 2012;95–101.

11 Wijaya TK, Vasirani M, Humeau S, Aberer K. Cluster-based aggregate forecasting for residential electricity demand using smart meter data. *2015 IEEE International Conference on Big Data*. 2015;879–887.

12 Hernández L, Baladrón C, Aguiar J, Carro B, Sánchez-Esguevillas A. Classification and clustering of electricity demand patterns in industrial parks. *Energies*. 2012;**5**(12):5215–5228.

13 Wang Y, Chen Q, Kang C, Xia Q. Clustering of electricity consumption behavior dynamics towards big data applications. *IEEE Transaction on Smart Grid*. 2016;**7**(5):2437–2447.

14 Maksood FZ, Achuthan G. Sustainability in Oman: energy consumption forecasting using R. *Indian J Sci Technol*. 2017;**10**(10):1–14.

15 Sibi S, Pushpalatha S. Demand forecasting of electricity consumption using dynamic clustering of time series data. *Int J Pure Appl Math*. 2018; **119**(15):525–533.

16 Gajowniczek K, Ząbkowski T. Simulation study on clustering approaches for short-term electricity forecasting. *Complexity*. 2018;**2018**:1–21.

17 Kore S, Khandekar S. Residential electricity demand forecasting using data mining. *Int J Eng Trends Technol*. 2017;**49**(1):27–32.

18 Auder B, Cugliari J, Goude Y, Poggi JM. Scalable clustering of individual electrical curves for profiling and bottom-up forecasting. *Energies*. 2018; **11**(7):1893.

19 Zhang GP. Time series forecasting using a hybrid ARIMA and neural network model. *Neurocomputing*. 2003;**50**:159–175.

20 Makridakis S, Weelwright S, Hyndman RJ. *Forecasting: Methods and Applications*. New York, NY: John Wiley & Sons; 1998.

21 Jakaša T, Andročec I, Sprčić P. Electricity price forecasting—ARIMA model approach. *2011 8th International Conference on the European Energy Market (EEM)*. 2011;222–225.

22 Hipel KW, Mcleod AI. *Time Series Modelling of Water Resources and Environmental Systems*. Amsterdam, The Netherlands: Elsevier;1994.

23 Wateo S, Churakham K, Intarasit A. Forecasting time series movement direction with hybrid methodology. *J Probab Stat*. 2017;**2017**:1–8.

24 Nepal B, Yamaha M, Sahashi H, Yokoe A. Analysis of building electricity use pattern using K-means clustering algorithm by determination of better initial centroids and number of clusters. *Energies*. 2019;**12**(12):2451.

25 Chen C, Twycross J, Garibaldi JM. A new accuracy measure based on bounded relative error for the time series forecasting. *PLoS ONE*. 2017;**12**(3): e0174202.

26 Snipes M, Taylor DC. Model selection and Akaike information criteria: an example from wine ratings and prices. *Wine Economics Policy*. 2014;**3**(1): 3–9.

27 Nepal B, Yamaha M, Sahashi H. Energy conservation in university buildings by energy pattern analysis using clustering technique. *Int J Energy Prod Manag*. 2019;**4**(2):158–167.

How to cite this article: Nepal B, Yamaha M, Yokoe A, Yamaji T. Electricity load forecasting using clustering and ARIMA model for energy management in buildings. *Jpn Archit Rev*. 2020;3:62–76. https://doi.org/10.1002/2475-8876.12135